UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

Semantic Similarity Across Biomedical Ontologies

DOUTORAMENTO EM INFORMÁTICA
ESPECIALIDADE BIOINFORMÁTICA

João Diogo Silva Ferreira

Tese orientada por:
Prof. Dr. Francisco José Moreira Couto

Documento especialmente elaborado para a obtenção do grau de doutor

2016

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

Semantic Similarity Across Biomedical Ontologies

DOUTORAMENTO EM INFORMÁTICA
ESPECIALIDADE BIOINFORMÁTICA

João Diogo Silva Ferreira

Tese orientada por:
Prof. Dr. Francisco José Moreira Couto

Júri:
Presidente:
Prof. Dr. Nuno Fuentecilla Maia Ferreira Neves
Vogais:
Prof. Dra. Maria Cristina de Carvalho Alves Ribeiro
Prof. Dr. José Luís Guimarães Oliveira
Prof. Dr. Nuno Manuel de Carvalho Ferreira Guimarães
Prof. Dr. Francisco José Moreira Couto (orientador)
Prof. Dr. Manuel João Caneira Monteiro da Fonseca
Prof. Dr. Francisco Cipriano da Cinha Martins

Financiamento: Fundação para a Ciência e Tecnologia

Documento especialmente elaborado para a obtenção do grau de doutor

2016

I wish to dedicate this thesis not to the usual suspects
(not my wife or parents, my friends or teachers)
but to those who I have had my mind on for
a long time, the ones who will benefit the
most from scientific discovery and from the
vast accomplishments achieved today.
I dedicate it, rather, to the ones who
will linger after us and so reap
the benefits of our hard work.

To my future children!

Acknowledgements

Foremost, I would like to express my gratitude to my advisor, Prof. Dr Francisco M. Couto, whose expertise and undeniably scientifically oriented mind added considerably to my PhD. I appreciate his vast horizons and skills in many scientific areas (*e. g.* semantic similarity, biomedical ontologies, Bioinformatics in general, just to list the ones more closely related to my research efforts), and his assistance in the extensive periods where writing was the major undertaking of my work (particularly the several manuscripts we jointly submitted for publication, as well as this document). Most of all, I look up to his "relaxed" yet serious attitude towards research, which always positively surprised me. Sometimes, research feels like a herculean task that we need to carry out in order to see some results, but more often it looks like a small train ride where we simply appreciate the landscape of knowledge to which we ourselves contribute.

My PhD companions, which are no longer students but Doctors, also deserve my gratitude: in no particular order, I thank Ana Teixeira, Catia M. Machado, Catia Pesquita, Daniel Faria, Hugo Bastos and Tiago Grego, with whom I collaborated in some way or another, and whom I could always count on to tag along on a brainstorming session over coffee. A further acknowledgement goes to André Lamúrias, with whom I collaborated towards the end of my PhD.

I recognize that this research would not have been possible without the financial assistance of the Fundação para a Ciência e Tecnologia, and express my gratitude to that agency for funding my PhD scholarship SFRH/BD/69345/2010.

I would also like to thank my parents for the infinite support they have been providing throughout my entire life, from the very first breath I took. The excellent upbringing you provided was the exact amount of motivation I needed to pursue the career in science that I now so dearly esteem. Thank you for always being there for me, for always helping me succeed. I also want to thank my brothers, André and Simão, for no other motive than that you are always present in my life. You are the best.

In particular, I must acknowledge my wife and best friend, Sandra, who stood by me every step of the way, who had to endure my absence whenever I needed to attend some conference, and without whose love, encouragement and unconditional assistance I would not have been able to finish this work. The biggest of my smiles is yours.

Lisboa, January 6$^{\text{th}}$, 2016

João Ferreira

Resumo

As práticas que levam à descoberta científica estão geralmente associadas à necessidade de comparar entidades relevantes para essa mesma descoberta. Por exemplo, em medicina a comparação de um novo caso clínico com uma base de dados de casos antigos pode ajudar uma equipa médica a acelerar o processo de diagnóstico ou mesmo de tratamento; em investigação laboratorial, a semelhança na estrutura molecular de compostos químicos é útil na pesquisa de novos fármacos.

Apesar da necessidade de comparar as entidades anteriormente exemplificadas, é difícil encontrar métodos automáticos que sejam reprodutíveis, generalizáveis, e que consigam lidar com a diversidade destes dados. Na verdade, qualquer método automático tem obrigatoriamente de se basear numa representação objetiva e computacionalmente tratável dos dados, *e. g.* objetos matemáticos como vetores ou sequências de caracteres. No entanto, a comparação de tais objetos é independente do contexto, *i. e.* estes algoritmos de comparação transformam os dados sem qualquer conhecimento do seu significado. Além disso, este tipo de representação elimina por vezes informação relevante acerca das entidades. Exemplos de soluções pontuais para alguns tipos de dados incluem:

- Comparação de compostos químicos através da sua estrutura molecular representada através de um grafo, onde os nós representam átomos e os arcos representam ligações químicas. Esta comparação pressupõe que uma semelhança na estrutura implica uma semelhança na atividade química dos compostos, uma relação que nem sempre é válida.
- Comparação de proteínas através da sua sequência de aminoácidos, o que como anteriormente nem sempre é válido uma vez que sequências parecidas não correspondem sempre a funções parecidas.

Por outro lado, há entidades que não são trivialmente representadas de forma matemática. Por exemplo, como comparar dois casos clínicos? Uma das formas de ultrapassar esta dificuldade é representar o próprio *significado* num formato que possa ser interpretado por computadores, uma prática conhecida como representação do conhecimento.

A representação do conhecimento é uma área de investigação que tem como objetivo definir formas de tornar o conhecimento manipulável por computadores, tal como é manipulado por seres humanos, o que permite aplicar sobre ele algoritmos de raciocínio automático. O princípio

mais útil desta área para o meu trabalho é que o conhecimento pode ser representado sob a forma de *ontologias*, que são definidas, de forma simplista, como um conjunto i) de conceitos relativos a um domínio do conhecimento, e ii) das relações entre eles. Por exemplo, uma ontologia que representa conhecimento anatómico define que Coração é um tipo de Órgão, e que Fémur é um Osso que *articula-com* a Tíbia. As ontologias fornecem assim um *significado* aos conceitos, não de forma explícita (como acontece por exemplo num dicionário, onde o significado é descrito em texto), mas de uma forma implícita, que emerge das relações definidas entre eles.

As ontologias permitem quantificar objetivamente o grau de semelhança entre os conceitos nelas contidos, através da comparação dos seus significados. Esta prática, conhecida como *semelhança semântica*, permite determinar, por exemplo, que Braço é mais semelhante a Perna do que a Coração, uma vez que Braço e Perna representam tipos de Membro, enquanto que Braço e Coração partilham entre eles apenas o conceito genérico Parte do corpo.

Além da comparação entre conceitos, as ontologias permitem a comparação de entidades que estejam *anotadas* com esses conceitos. No contexto deste documento, anotar uma entidade consiste em associar a essa entidade informação computacionalmente tratável que a descreva, geralmente através de conceitos de ontologias. Uma anotação não é mais do que uma descrição objetiva de um facto: "esta entidade está relacionada com este conceito". Por exemplo, é prática comum anotar as proteínas de uma base de dados com conceitos que representam as suas funções. Ao comparar as anotações de duas proteínas estamos efetivamente a comparar as duas proteínas; torna-se assim possível utilizar o próprio significado biológico das proteínas para as comparar, não sendo necessário recorrer a representações mais simplistas, como a sequência de aminoácidos.

Um dos aspetos da prática de semelhança semântica pouco investigados até ao momento é a questão da multidisciplinaridade dos dados. Em particular, na informática biomédica, os dados existentes são geralmente descritos com recurso a múltiplos domínios de conhecimento. Por exemplo, um caso clínico pode estar anotado com conceitos relativos aos sintomas, aos resultados de análises ao sangue, aos medicamentos prescritos, ou até a conceitos menos óbvios como os locais anteriormente visitados pelo paciente ou as suas condições sócio-económicas. Todos estes aspetos podem influenciar o diagnóstico e o tratamento escolhido, sendo portanto essenciais para um cálculo preciso da semelhança entre casos clínicos.

Este documento reporta a minha investigação na área da semelhança semântica no que respeita à sua aplicação em contexto multidisciplinar. Até à data, não existem trabalhos científicos publicados neste campo de investigação, sendo este o primeiro a surgir, não só na comunidade da informática biomédica como também no resto da comunidade científica. Eu proponho duas abordagens para comparar entidades multidisciplinares:

1. Na abordagem *agregativa*, todos os domínios de anotação são usados de forma isolada uns dos outros para calcular vários valores de semelhança unidisciplinar com um algoritmo pré-existente (*e. g.* os sintomas de um caso clínico são comparados com os sintomas do outro caso clínico, depois os medicamentos de um caso com os

medicamentos do outro, depois as condições sócio-económicas, *etc.*). Os valores são por fim agregados matematicamente, *e. g.* através da média.

2. Na abordagem *integrativa*, todas as ontologias relevantes são unificadas numa grande ontologia multidisciplinar, sendo em seguida aplicado um algoritmo de semelhança semântica pré-existente para comparar o conjunto de todas as anotações de uma entidade com o conjunto de todas as anotações da outra entidade.

Ambas as abordagens se baseiam na existência de um algoritmo capaz de comparar dois conjuntos de conceitos provenientes de uma só ontologia (*algoritmo de ontologia única*). Esta escolha baseia-se no facto de já existirem várias medidas capazes de executar essa tarefa, amplamente estudadas e aplicadas em vários casos. Além disso, as duas abordagens propostas são independentes da medida pré-existente utilizada, sendo portanto possível utilizar uma medida apropriada para o contexto em questão.

A metodologia seguida para provar que as medidas de semelhança multidisciplinar são eficazes no seu objetivo consistiu essencialmente em cinco passos:

1. recolher dados multidisciplinares;
2. validar as duas abordagens acima descritas aplicando-as aos dados recolhidos;
3. sistematizar os métodos de validação de semelhança semântica;
4. propor melhorias às medidas de semelhança de ontologia única; e
5. criar *software* para calcular a semelhança semântica de forma reprodutível.

Ao longo do meu doutoramento, recolhi três conjuntos de dados multidisciplinares anotados com conceitos de várias ontologias (passo 1): *i*) um conjunto de artigos científicos da área da epidemiologia anotados com conceitos como doenças, sintomas, vacinas, modos de transmissão, *etc.*; *ii*) um conjunto de vias metabólicas anotadas com compostos químicos, enzimas, doenças associadas a erros na via metabólica e drogas que afetam o funcionamento das vias; e *iii*) um conjunto de modelos matemáticos de sistemas biológicos, anotados com compostos químicos, enzimas, entidades anatómicas e fenótipos.

Para validar as abordagens multidisciplinares (passo 2), segui três estratégias de validação distintas: *i*) prever novas anotações com base nas já existentes, *ii*) classificar automaticamente os dados com base nas suas anotações, e *iii*) comparar os valores de semelhança obtidos automaticamente com valores de semelhança atribuídos manualmente por especialistas. Em cada uma destas estratégias, calculei uma métrica de desempenho que permite determinar a validade das medidas de semelhança semântica.

Os resultados obtidos com as três estratégias de validação mostram empiricamente que as medidas de semelhança multidisciplinares são de facto eficazes. Por comparação do desempenho atingido por cada abordagem multidisciplinar (agregativa e integrativa), conclui-se que a abordagem integrativa é em geral superior à abordagem agregativa. Este resultado corresponde ao que era esperado, uma vez que a abordagem integrativa tem acesso a mais informação, pois

utiliza relações existentes entre conceitos de duas ontologias diferentes e portanto atinge valores de semelhança mais precisos.

Comparei ainda o desempenho das abordagens multidisciplinares com o desempenho obtido com as medidas de semelhança de ontologia única. Apenas em alguns casos excecionais as abordagens multidisciplinares não foram superiores às medidas de ontologia única (esta situação ocorre essencialmente quando se utilizam anotações de uma ontologia que já representa, por si mesma, vários domínios de conhecimento, e portanto a adição de novos domínios de anotação não melhora o resultado).

Outras contribuições importantes foram atingidas com este trabalho. Nomeadamente:

- Estabeleci uma hierarquia de estratégias de validação para utilizar no desenvolvimento de semelhança semântica, a qual permite organizar os vários métodos de acordo com o tipo de aplicações onde podem ser usados (passo 3).
- Propus novas medidas de semelhança de ontologia única que exploram uma maior quantidade da informação representada nas ontologias, aumentando não só o leque de algoritmos disponíveis mas também o seu desempenho (passo 4).
- Desenvolvi uma infraestrutura de *software* que permite obter resultados de semelhança semântica não só de forma rápida mas reprodutível, sendo extensível, ou seja, permite que outros investigadores facilmente implementem os seus algoritmos de semelhança semântica (passo 5).

Com o aumento da quantidade de conhecimento que nós humanos vamos construindo e ao qual vamos tendo acesso, é a simbiose entre investigadores e métodos automáticos que permite gerir o conhecimento de forma eficiente. Só assim poderemos, como comunidade, garantir a descoberta de nova informação e assegurar que ela pode ser utilizada no futuro. A semelhança semântica é apenas um dos aspetos desta automatização.

A minha contribuição, como quase todas em ciência, permite vislumbrar apenas um pouco além da fronteira que engloba o conhecimento humano, mas juntamente com as descobertas de milhões de outros cientistas, está a construir e a melhorar o conhecimento que temos do nosso mundo e de nós próprios. Assim, considero que o meu trabalho é um pequeno mas robusto passo em direção ao futuro da Ciência.

Palavras-chave Semelhança semântica multidisciplinar, Ontologias biomédicas, Anotação de entidades biomédicas, Web Ontology Language, Web semântica, Validação de semelhança semântica, Multidisciplinaridade dos dados biomédicos

Abstract

The need to compare complex entities is relevant in all the areas of science. In medicine, for example, comparing a clinical case to a database of previous cases can be extremely helpful when trying to diagnose a disease or deciding the most appropriate treatment for a patient.

Recent developments in knowledge representation, in particular the creation of the Web Ontology Language (OWL), have lead to a rise in the amount of knowledge that is being stored in *ontologies*, which represent, in machine-readable format, the known facts about reality. With the help of ontologies, statements like "Influenza is an Infectious disease" can be processed by computers, which, in turn, can be used to create new knowledge. In particular, *semantic similarity* has emerged to explore these ontologies as a way to compare entities annotated with the ontology concepts.

Semantic similarity has been extensively studied in the last decade, but some problems still persist. While there are algorithms to compare entities annotated with concepts from the same ontology, the possible ways to use *more than one ontology* are still in an early phase of study. For example, comparing a metabolic pathway using both the associated molecular functions and the metabolites converted in the pathway should, in principle, yield a higher precision than would be achieved with methodologies that rely on either one of the two domains independently. Comparing concepts from *different domains* and entities annotated with concepts from different domains is yet an unexplored area, but necessary to tackle multidisciplinary biomedical resources, *e. g.* to compare two clinical cases, the relationships between symptoms, diseases, blood screening results, *etc.* should provide a more insightful and precise value of similarity.

In this document, I explain the basic concepts needed to understand the problem of semantic similarity, how it is being solved, and how I propose to extend this notion so that it can be applied to more than one ontology and, more significantly, to more than one domain of knowledge.

Keywords Multi-domain semantic similarity, Biomedical ontologies, Biomedical annotated entities, Web Ontology Language, Semantic web, Semantic similarity validation, Multidisciplinarity of biomedical data

Contents

List of Figures

List of Tables

List of Listings

PART I

Background

> Imagination is the beginning of creation. You imagine what you desire, you will what you imagine and at last you create what you will.
>
> — George Bernard Shaw

Chapter 1

Introduction

The process of finding similar and related concepts is one of the most characteristic activities of human nature and, specifically, of scientific research. Indeed, the categorisation of known concepts (*i. e.* the proper distribution of the concepts in manageable categories, where each category contains only concepts that are similar to each other) introduces an abstraction layer over the reality that enables a more focussed reasoning over experimental results and empirical observations. The importance of categorisation is tightly coupled with the amount of concepts that one must deal with: as the amount of collective human knowledge increases, so does the needs for good categorisation that abstracts away the less useful details of reality, thereby increasing its manageability.

Therefore, similarity and relatedness measures are, without a doubt, necessary assets for the advancement of science. As we will see in this document, I argue that *semantic* measures are, in fact, one of the most useful ways to achieve similarity and relatedness values for today's scientific resources, which are now sufficiently ripe for use in research, with particularly applicable results in the context of biomedical informatics.

The findings of this thesis focus primarily in the biomedical domain, given that the level of commitment in the biomedical informatics community to develop automatic systems to help their research (and ultimately contribute to the medical practice) is extremely high. For this reason, the whole document is written with a biomedical point of view. However, it is important to notice that the results that I will describe later and the contributions that stem from the work that I carried out can be generalised into other areas of research with minimal effort.

1.1 Motivation

It is an undeniable and undisputed fact of scientific research that the amount of knowledge and data published each year is increasing at an exponential pace [LI10]. This behaviour is true not only across areas of research but also across types of publication: it can be observed in the number of academic papers as well as in the sizes of databases, with staggering examples in protein databases [*e. g.* The10] or chemical databases [Wil08]. Even clinical information can

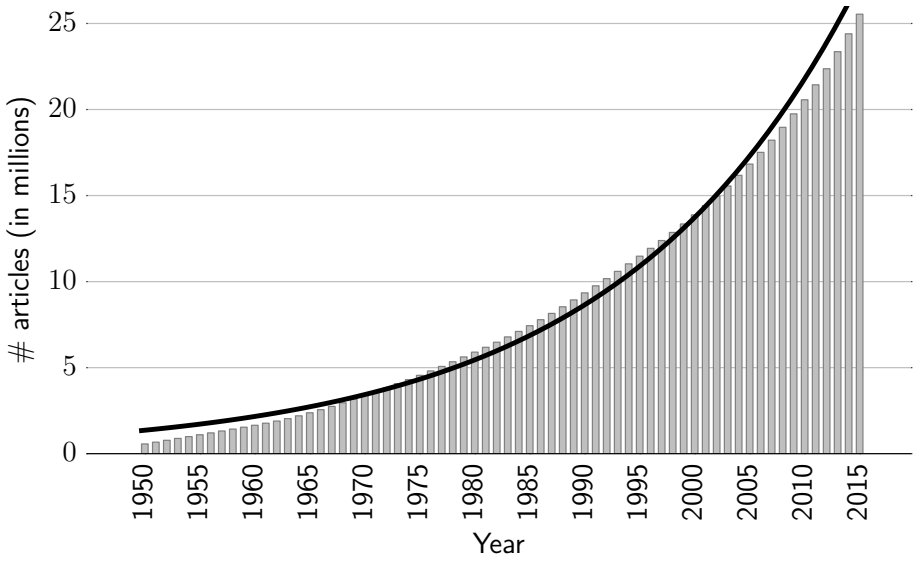


Figure 1.1 – Yearly size of MEDLINE from 1950 to 2015. This plot shows that the number of bibliographic entries in MEDLINE has increased exponentially. The bars represent the cumulative number of articles indexed in this database each year, while the bold line is an exponential fit to the data. This rate of growth corresponds, on average, to a doubling in the amount of articles every 15.0 years. Data retrieved from an actual search for publications between 1950 and 2015, in http://www.ncbi.nlm.nih.gov/pubmed. This image is meant as an illustration only, as there are many scientific results that are not published in MEDLINE-indexed journals.

be collected (subject to ethic guidelines, given the need for confidentiality in such records) and used in research, a practice called translational medicine [Nal06; Weh08]. **Figure 1.1** illustrates this facet of scientific discovery by plotting the amount of articles contained in MEDLINE (a bibliographic database for the life sciences) against time, which shows an approximately exponential trend. The same behaviour can be observed in other databases of biomedical information, such as the one in **Figure 1.2**.

With this exponential increase, two major problems have arisen. First, it has become impossible for researchers to fully read and interpret all the new information that becomes available each day. Second, and more important in scientific research, data published by different authors is often released in different formats, with different assumptions on what the meaning of each particular datum is. This makes it difficult, and in some cases impossible, to properly integrate all this information under a central knowledge repository without a lot of effort on the part of the data owners.

Surprisingly, these two problems are related, and solving one will help solve the other. Specifically, the impracticality of manually reading papers has created the need to develop automatic systems able to read textual documents, to appropriately parse and interpret them, and finally to draw conclusions based on them (for example by generating an automatic summary).

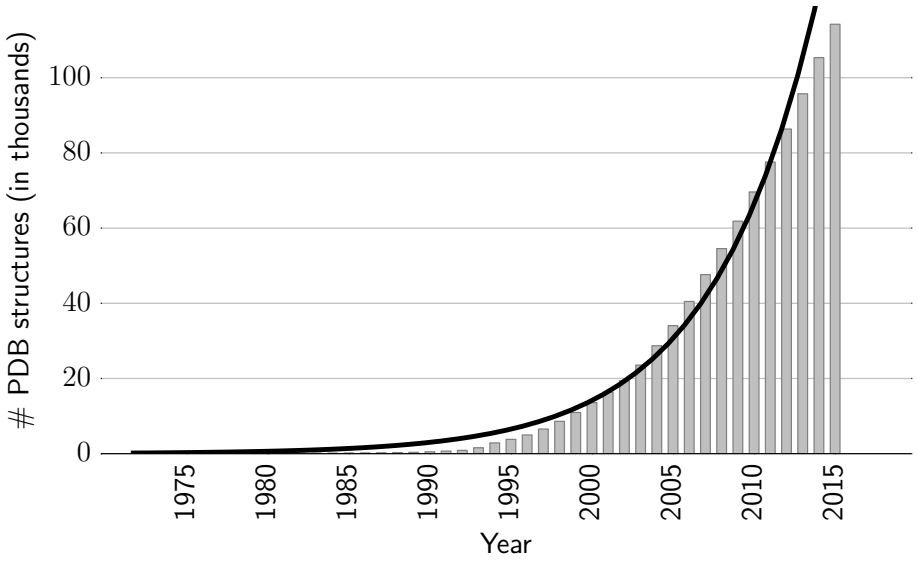


Figure 1.2 – Number of 3-dimensional protein structures in PDB from 1975 to 2015. This plot shows that the number of protein structures stored in the Protein Data Bank has increased almost exponentially since the database has been created. The bars represent the total number of structures in this database, while the bold line is an exponential fit to the data. This rate of growth corresponds, on average, to a doubling in the size of the database every 4.5 years. Data from http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total.

To allow such a system to function, it must be able to understand the *meaning* underlying the concepts referred to in the text, which, for example, includes understanding that the heart is responsible for pumping blood, that muscles are attached to bones through tendons, or that infections can cause fever. This, in turn, requires that knowledge be encoded in a machine-readable format, which must be precise, formal, and comprehensive, enabling computers to perform reasoning. Only that allows a computer to interpret text. On the other hand, knowledge that is described in such a format, using standard representations that everyone agrees with, can also be used by computers and, ultimately, be regarded as interoperable. This solves the second issue: if an automatic system is able to parse data from multiple sources in a logical format, it can use them uniformly as if they had the same provenance.

There are, today, many organisations dedicated to the production of formal standards for knowledge representation as well as *reference knowledge artefacts* (called ontologies throughout this document), predominantly in the biomedical informatics community. For example, the Gene Ontology (GO), first published in 2000 [Ash00], is an attempt to "address the need for consistent descriptions of gene products in different databases" [The12]. With the maturation of the standards for knowledge representation and the semantic web (see Sections 2.1 "Knowledge representation", 2.2 "Ontologies" and 2.4 "Semantic web" for a summary of the importance of these concepts to the field of biomedical informatics), GO has grown into a mature digital artefact

that represents properties of proteins such as molecular function and cellular localization.

To explore the wealth of information that is stored in biomedical data, it is vital that ontology developers and users be provided with tools that can properly manage and handle the data; one ubiquitous requirement in science is the ability to estimate the degree of similarity or relatedness between the various ontology concepts [Vis11] and, by extension, between multidisciplinary entities. For instance, similarity between proteins is often associated with one or more functions being shared among them [Alt90]; in chemistry, similarity of molecular structure correlates with similar biological role [HF64; KAM94]; in medicine, a high degree of similarity between two clinical cases is a strong argument towards a similar diagnosis [Swe74].

One of the technologies enabled by the use of ontologies is indeed the calculation of similarity between the concepts they represent, a technique known as "semantic similarity" [Res95; Lor03; Pes09b]. This technique can be used to compare concepts within one ontology as well as entities annotated with those concepts. For example, proteins annotated with GO functions can be compared based on the semantic similarity of those functions. While traditional automatic systems compared proteins by their sequence, using methods such as BLAST [Alt97], semantic annotations provide a mechanism to compare proteins by their functions [Lor03]. Advantages of this include the fact that some proteins that are known to have similar functions have different sequences, or vice-versa. Thus, semantic similarity can explicitly explore information about entities (in this case proteins) to more accurately compare them.

Semantic similarity has been applied to various domains:

- between proteins annotated with GO concepts describing their molecular functions [Lor03; LD06];
- between metabolic pathways annotated with their enzymes [CSV05] or chemical compounds [Gre10]; and
- between diseases annotated with biological processes [SLA10].

Real world problems whose solution incorporates semantic similarity include the prediction of *i*) the probability of a certain disease given a set of symptoms [Köh09], *ii*) the cellular localization of proteins given their annotations [LD06], *iii*) the function of proteins [Pes08], and *iv*) the chemical properties of small metabolites [FC10].

One of the remaining issues in this area is related to the fact that ontologies are often developed with a single domain of reality in mind. GO represents knowledge associated with proteins, CHEBI (Chemical Entities of Biological Interest) represents knowledge associated with biologically relevant chemical substances [Deg08], FMA (Foundational Model of Anatomy) represents human anatomy [RM03], *etc.* But knowledge is multidisciplinary, with some concepts from one domain being often intertwined with concepts from another domain. This is particularly true in the biomedical field, which is so vast that it is partitioned in several distinct but related disciplines. Clinical cases, for example, may include information on the symptoms, blood screen results, and drugs the patient is currently taking; even less obvious concepts, such as the previously

visited places and the economical conditions of the patient can prove useful in tracing a diagnosis. Models of metabolic pathways refer to the reactions, the enzymes, the chemical metabolites the cellular components involved in the pathway, *etc.* When accuracy is necessary (and it frequently is in the biomedical domain), multidisciplinarity naturally arises.

While single-ontology semantic similarity has been extensively studied in the last two decades, current algorithms are unable to adequately compare multidisciplinary entities. To handle multidisciplinarity, I propose that it is possible to harness the advantages of the existing single-ontology measures to allow their use in this context, by *lifting* them from the single-ontology constraints. For this purpose, I study the following two approaches:

1. Use single-ontology measures to compare concepts from the same domain in the two entities with single-ontology measures (*i. e.* compare the chemical reactions of one entity with the chemical reactions of the other entity, then cellular components with cellular components, symptoms with symptoms, *etc.*) and then combine the various results in a single value by means of an aggregating function such as the average. I call this the *aggregative* approach.
2. Use the inherent expressiveness of ontologies to integrate all knowledge in a single multidisciplinary knowledge base. Instead of calculating a value for each domain, this approach exploits the inter-domain links that exist between the ontologies to calculate multi-domain similarity and relatedness. For example, the relationship between skin and rash can be used to link together an ontology of anatomy and one of symptoms. I call this the *integrative* approach.

1.2 Objective

The theoretical objective of my PhD was to prove the following thesis:

> “Multi-domain semantic similarity measures can be constructed by lifting single-ontology measures according to the two approaches defined above (the aggregative and integrative approach), thus enabling semantic similarity on multidisciplinary entities.”

This statement is the driving force behind all the research efforts related to my work. In particular, I expect *i*) that the comparison of multidisciplinary entities will be more effective when using a multi-domain measure rather than a single-ontology measure applied only to one domain, and *ii*) that the integrative approach will generally be more effective than the aggregative approach, as it has access to more information.

Effectiveness is an abstract concept that can have several interpretations depending on the context in which it is applied. In a medical context, a measure is effective if it can, for example, predict a disease from the clinical notes associated with a patient; in pharmacology, a measure is effective if it can be used to diminish the costs of drug tests by pre-emptively filtering potential

drugs, thereby reducing the number of necessary trials. The proposed hypothesis, however, is orthogonal to the measure of effectiveness that one uses: irrespective of the way effectiveness is calculated, multi-domain measures perform well.

The practical and more fundamental objective of this thesis is, therefore, the creation of both *i*) a semantic similarity framework that is able to deal with multidisciplinary entities, and *ii*) semantic similarity measures that compare these entities in a way that reflects their actual meaning. For the first part, I will develop a system that is able to use existing single-ontology measures and lift them to multi-domain measures. For the second part, the focus will fall on finding use cases where multi-domain semantic similarity is needed, such as the detection of similar biochemical pathways or similar epidemiological resources. These datasets will be used to evaluate the effectiveness of multi-domain measures.

1.3 Methodology

To achieve the main objective of this thesis, I had to fulfil five separate tasks.

1. Study how validation is done in the field of semantic similarity Biomedical research in this field has been generating innovative and useful measures of similarity since 2003 and applying them to several distinct problems; however, validation is still being done in a relatively *ad hoc* way, where each proposed measure is validated with a different method without much relation to previous ones. While some steps have been followed to mitigate this problem, a true systematization of validation strategies is still lacking, and as such one of the tasks of my work will be to determine to what extent this problem can be alleviated.

2. Enhance current similarity and relatedness measures With the recent advance in the research of knowledge representation, ontologies are becoming increasingly richer and more expressive, and tools are being developed to handle this expressiveness. However, semantic similarity is not following this trend. For example, most algorithms are agnostic to the ideas of formal axioms, and as such are unable to use facts like the ones expressed with disjoint axioms (*i. e.* there is no thing that is both a Square and a Circle) or other logic axioms. I believe that exploring in more detail the formal logic aspect of ontologies will yield measures of similarity that better reflect the structure of the ontology and the relationships between its concepts.

3. Collect multidisciplinary datasets To test the measures of similarity and relatedness developed in this thesis, it will be necessary to collect multidisciplinary data annotated with concepts from various ontologies, which will allow the use of the aggregative and integrative approaches. Another important aspect of this task is the possibility to use the data collected and the help of experts to create gold-standards that can be used to validate the measures of similarity.

4. Validate the multi-domain measures This task will finalise the proof of the proposed thesis by finding evidence that supports it. Validation of multi-domain semantic similarity measures can be done in several ways. For example, by comparing the automatically assigned similarity values with the ones assigned by experts in the gold-standards created in the previous task, or by using it in classification problems and quantifying the difference in performance between the single-ontology measures and the multi-domain approaches.

5. Develop semantic similarity software Given the increasing number of ontologies, the amount of multidisciplinary data being published, and the growing standardization efforts in knowledge representation, it is more important than ever to develop the right tools to enable semantic similarity calculations in a reproducible way. As part of my contribution to this field, I will develop extensible software that will, on the one hand, allow developers to implement their semantic similarity measures under a common framework, and, on the other, provide users of semantic similarity (*e. g.* online data repositories) a way to quickly calculate similarity between concepts or between annotated entities. This technical task will assist the previous task by allowing quick calculation of semantic similarity.

1.4 Contributions

The contributions of this work can be summarised in terms of major and minor contributions. The five main contributions are aligned to the points delineated above:

1. a hierarchy of validation strategies that can be used to classify research in semantic similarity according to the way the measures have been validated;
2. an implementation of a single-ontology semantic relatedness measure, which can be generalised to the multi-domain context [FC11], and of a single-ontology semantic similarity measure that can deal with disjointness axioms [FHC13];
3. the compilation of three multidisciplinary datasets, in the areas of epidemiology, metabolic pathways, and biochemical models, annotated with ontology concepts, onto which the multi-domain measures of similarity and relatedness operate;
4. a validation of the two multi-domain approaches, by testing them on the multi-disciplinary datasets and verifying that they outperform single-ontology measures; and
5. the development of `OWLtoSQL` and `MOSSy`, two programs that work in tandem to provide easy semantic similarity calculations and which, in fact, already provide the implementations of the two multi-domain approaches.

In the course of my work, I have additionally contributed to the epidemiology and geographical domains. The first of these minor contributions was the creation of a network of ontologies that are relevant in the domain of epidemiology and allow the formal categorization and annotation of

epidemiological resources with ontology concepts, giving them semantic information that can be analysed by techniques like semantic similarity [Fer12]. I have also contributed to an alignment between a geographical ontology of the Portuguese territory (Geo-Net-PT) and another ontology encoding the geo-political divisions of the world [Fer10]. Furthermore, I used semantic similarity in this domain to create a disambiguation algorithm that maps geographical names in text to the correct concept in Geo-Net-PT [Bat12]. Finally, I have helped develop a text-mining system for the chemical domain that uses semantic similarity to validate its results [LFC15].

A brief summary of my main contributions can be examined in Section 9.1 "Summary of contributions", and is complemented in Appendix B "Auxiliary projects" with a small description of my minor contributions.

1.5 A word on terminology and notation

Throughout this document, I will make numerous references to terms that are essential to describe the field of semantic similarity. Chapter 2 "Concepts" will explain most of these terms, both to introduce the reader to these notions and to standardise the terminology, thus allowing a more thorough understanding of the document. To further assist the reader, the following typographical notation is used:

- a blackboard font is used for ontology acronyms (*e. g.* $\mathbb{GO}$, $\mathbb{CHEBI}$);
- a sans-serif font is used to refer to concepts, always starting with a capital letter (*e. g.* Head, ATP binding); and
- *italic shape* is used for relationships between concepts, always in lower case and with spaces translated to hyphens (*e. g. part-of*).

1.6 Structure of this document

This document is organised in four parts.

The first part deals with the contextualization of the problem underlying the proposed hypothesis. Chapter 2 defines and explains the basic notions needed to understand the problem itself, and Chapter 3 surveys the *state of the art* with respect to how semantic similarity has been conducted both in the single-ontology and multiple-ontology contexts.

The second part accounts for my contributions. It contains five chapters, in parallel to the five points delineated in the methodology. Namely, Chapter 4 describes how the hierarchy of validation approaches was constructed, Chapter 5 outlines the enhancements that I propose for improving single-ontology semantic similarity, Chapter 6 presents three multidisciplinary datasets collected to test the hypothesis, and Chapter 7 formally defines the two multi-domain semantic similarity approaches and demonstrates their performance on the multidisciplinary datasets, thus establishing the validity of the proposed hypothesis. Chapter 8, although an indispensable

part of the document, describes not direct scientific research but rather the technical aspects necessary for the execution of this methodology, by characterising the software that I developed to perform semantic similarity calculations.

The third part is composed of Chapter 9, which enumerates some conclusions, limitations of my contributions and potential future work.

The last part deals with the appendixes, where I explain some of the details of the work in more detail than was possible in the main document, including Appendix A, which contains a list of relevant ontologies used throughout my work and Appendix B, which describes three research efforts where I participated that are related (if only tangentially) to my work. Finally, the document ends with a list of references.

Chapter 2

Concepts

In this chapter, I will present to the reader a set of concepts that are necessary to fully understand the scope of this document and its implications for the future of scientific research.

2.1 Knowledge representation

As discussed in the introduction, the collective knowledge of mankind is ever increasing, and scientific knowledge is no exception. Measuring this growth is not easy, but the truth of this statement is often illustrated by pictures such as the one in **Figure 1.1**, which plots the number of articles indexed by MEDLINE through time. This increase seems to be exponential, which can be stated in other words: the amount of knowledge produced depends on the amount of knowledge that exists. The more we know, collectively, as a society, the more we can discover.

With this increase, managing, processing and using the total amount of knowledge becomes more difficult to do. This is where the power of computers can be harnessed to help us in the endeavour of knowledge discovery. The difficulty with this is that knowledge is not directly machine readable. Indeed, established facts have been traditionally published in plain text, which enables humans to understand them; however, natural language processing techniques are not yet fully capable of converting scientific text into *actionable* formats (*e. g.* formats that allow automatic reasoning). Therefore, to enable the application of computerised processing power to knowledge manipulation, it is essential that we find ways to represent knowledge in a machine readable format, which is the subject of Knowledge Representation (KR).

The goal of KR is to find ways to give machines the means to deal with information the same way that humans do, which will ultimately allow them to reason over data and create new knowledge, or at least assist humans to do so. Under this point of view, KR can be (grossly) reduced to two related tasks: *i*) establish the right formats for representing knowledge, and *ii*) specify and implement reasoning capabilities that exploit the knowledge thus represented. In this thesis, I will lean on both aspects of KR: first, I will focus on the representation aspect of KR, which provides the information needed to implement similarity measures; second, semantic similarity itself enables reasoning over data (for example, proteins of similar function often have

the same sub-cellular localization, and similarity of function can be used to infer this).

However, the subject of KR is vast, with roots in logic, psychology, and even mathematics. I will, therefore, only lightly touch these subject, and always indirectly (as fascinating as it may be, a full treatment of this subject is outside the scope of this document). For example, although I will use the notion of "Ontologies" (see next section) as the medium through which knowledge is represented, the full notion of logical formalisms will be mostly absent.

There are two kinds of KR-based reasoning: *i*) deductive reasoning, *i. e.* drawing specific conclusions based on general information (a true fact about animals can be used to deduce a true fact about humans, for example that both are living beings), and *ii*) inductive reasoning, *i. e.* drawing general conclusions based on specific data points (a known fact about a large set of mammals can be used to induce that the same fact is true about any mammal, for example after observing that dogs, lions, cats and giraffes have fur, we can induce that all mammals have fur) [Ove13, chap. 1]. Semantic similarity, being a tool that facilitates automated reasoning, can only be used to produce inductive arguments. However, inductive reasoning is not true reasoning, in the sense that it can reach wrong conclusions (the example above being an illustration of that), but it provides a starting point for further experimentation. For example, semantic similarity can be used to predict a set of probable (but not certain) functions for a given protein, which must then be tested in wet laboratory conditions. Under this context, semantic similarity can be used in techniques such as machine learning to produce new knowledge that is not logically derived from existing one but is instead induced from the starting data.

2.2 Ontologies

The term "ontology" was originally used by philosophers, meaning the study of reality, of what exists and how the existing things can be organised and subdivided based on their differences and similarities. The practice of categorising the reality in this manner is in fact an old one, going back to Aristotle (circa 350 BC), who tried to categorise all living things into a hierarchy based on the apparent complexity of their structures and functions, a *scala naturae*, or "ladder of life", as it was later called by Singer [Sin31].

This term has more recently been borrowed by computer science to mean a particular computational artefact (for example, a computer file, or a database) that contains *i*) a set of concepts belonging to a certain domain of knowledge, and *ii*) the ways these concepts relate to each other [Gru93]. One important aspect of computational ontologies is the notion that the ontology actually provides *semantics* (*i. e.* meaning) to the concepts it represents; however, the meaning is not described explicitly, as happens for example in dictionaries and glossaries, but rather *emerges* from the relationships between the concepts and the overall structure of the ontology. Consequently, in a real and useful way, ontologies are machine-readable representations not only of knowledge and facts, but also of the meaning of the concepts pertaining to a given domain of reality, and of the relationships between these concepts.

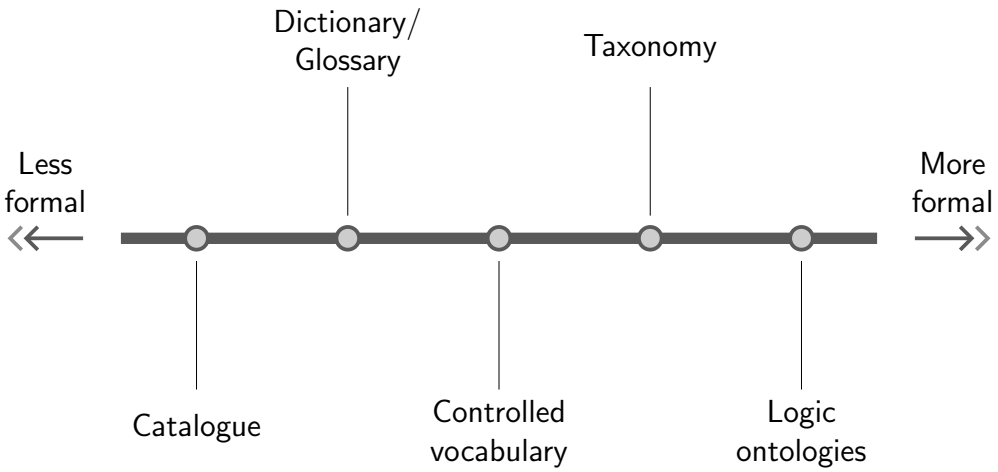


Figure 2.1 – The spectrum of ontology formality. This picture depicts the possible interpretations of what constitutes an ontology, arranged by formality levels. A *catalogue* is a list of terms, with no definitions or relations between themselves; a *dictionary* provides some textual definitions, as well as, possibly, synonyms; a *controlled vocabulary* defines a standard set of terms to be used in a particular context, along with textual definitions, synonyms, related concepts, *etc.*; a *taxonomy* arranges the terms is a hierarchy, without formally defining what it means for a concept to be classified under another concept; and a *logic ontology* provides machine-readable knowledge in a formal and precise way. Many other possible interpretations have been left out of this spectrum. Adapted from [McG02].

Within the computer science community, it seems there is no agreed-upon definition of what an ontology is [Gua98]. An oft cited definition is that an ontology is "an explicit specification of a conceptualization" [Gru93], but this vague and abstract description makes it difficult to properly visualise the true meaning of the word. In this line of thought, in fact, there are a number of potential information artefacts that can be regarded as "specifications" of domains of knowledge, where the main difference between them is their *formality*. **Figure 2.1** presents some possible artefacts that have been at one point in history, regarded as ontologies, arranged by formality levels. The more formal an ontology is, the more precise and expressive is the knowledge it represents.

This spectrum can be used to distinguish some of the ontologies used by the biomedical informatics community. While the most recent ontologies, including most of the ontologies used in this work, have been developed using formal systems (*i. e.* with the use of formal languages), some have still not been fully formalised. An example is the Medical Subject Headings (MeSH), which is a taxonomy of concepts that are related to one another by means of an underspecified relationship type. For example, Head is categorised under Body Regions, and Ear is categorised under Head, but while heads are body regions, ears are not heads; they are instead *parts* of the head. This illustrates the informality of MeSH: only one relationship type exists, but it is used to express different notions.

Contrast this with the far right end of the spectrum, occupied by fully formal ontologies.

In these, complex logic-based assertions can be made about the domain being represented in the ontology. For example, one can express the notions that *i*) Square and Circle are disjoint concepts (nothing can exist that is both a square and a circle); *ii*) Elephant is a subclass of Animal (all elephants are animals); and *iii*) a Finger is *part-of* some Hand.

In order to develop fully formal ontologies, with formal-logic constructions that allow us to assert those types of facts (called *axioms* in KR), several languages have been developed over the years, enabling knowledge engineers to formally represent the concepts of a domain and the relationships between those concepts. The current standard in KR is to use the Web Ontology Language (peculiarly, abbreviated as OWL). This language uses many first-order logic constructions to state facts about the concepts that are represented in the ontologies. OWL ontologies can be saved in files using different but equivalent formats, such as XML, Turtle or JSON. OWL semantics are specified by the World Wide Web Consortium, which defines how each construct should be interpreted and which logical conclusions can be deduced from them [MPG12; Mot12].

The most frequent construction in an ontology is the "class-subclass" relationship (variously called the *is-a*, "hypernymy" or "subsumption" relation): *e. g.* the concept Elephant is a "hyponym" of the concept Animal, since all elephants are animals (likewise, Animal is a "hypernym" of Elephant). Another common relationship between concepts is "meronymy", which is the relationship between the part and the whole, *e. g.* a Finger is part of some Hand. A significant portion of the knowledge represented in an ontology can be visualised as a graph, where nodes are concepts and edges are relationships (*e. g.* the class-subclass relationship, or the relationship between part and whole). The hypothetical ontology presented in **Figure 2.2** shows this parallel between ontologies and graphs. Each of the edges corresponds to one of the axioms of the ontology and, therefore, to an asserted fact. Usually, the class-subclass hierarchy is represented as a tree, while the other relationships are represented as general edges between the nodes.

In the biomedical domain, ontologies are often composed of concepts but not of actual instances of these concepts. For example, while the concept for Head exists in the human anatomy ontology, it being the machine-readable representation of all the human heads, the various instances of this concept (my head, the reader's head, and all human heads that ever existed and will ever exist) are not part of the ontology. In fact, since ontologies are abstractions over reality, they contain only facts that are true for all instances of a particular type. As such, they do not contain instances but instead represent concepts only.

As ontologies are inserted in a computer-science context, developing an ontology is in practice a two-sided task: on the one hand, it requires a logic background, since the formalisms of OWL are founded on first-order logic; on the other hand, it requires a background on the domain being represented in order for the ontology to be as accurate as possible. Beyond these inevitable prerequisites, building an ontology that aspires to be *the* standard representation of a scientific field of research demands a significant commitment to the best practices of ontology development. For example, ontologies should be reusable, the concepts should have textual

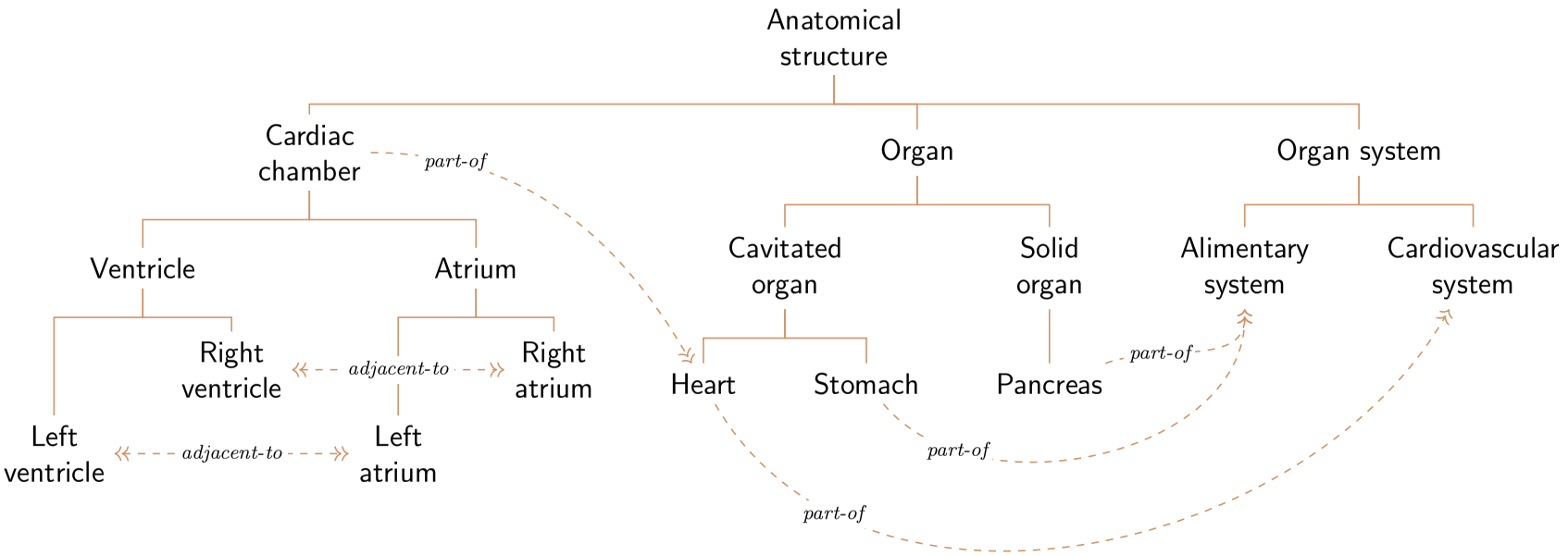


Figure 2.2 – A hypothetical ontology of human anatomical concepts. This hypothetical ontology contains several concepts related to the domain of human anatomy. It includes concepts that are related to one another by means of the class-subclass relationship (in solid lines) and other types of relationship (in dotted lines, edges labelled with the type of relationship). A note for the reader: this image will be referenced throughout this chapter and the next — take a moment to memorise that this is page number 17.

definitions for the benefit of its human users that are synchronised with the formal definitions (given by the axioms of the ontology that refer to the concepts), and the ontology should be kept updated in light of scientific advances, to guarantee its correctness [NM01].

Additionally, ontologies should be as interoperable as possible. There is a project within the biomedical informatics community, the OBO Foundry, which specifies a set of principles that are designed to increase the interoperability of the ontologies, such as orthogonality between ontologies and reuse of concepts form one ontology to the next [Smi07]. This ensures that each biomedical concept has a single representation and is, therefore, unambiguous.

2.3 Web Ontology Language

While research on semantic similarity should be agnostic to the languages used to express the ontologies, in practice, the existence of standards and community-driven recommendations means that the majority of ontologies are expressed using the same standards. In fact, the Web Ontology Language (OWL) is currently the language of choice to represent scientific knowledge, particularly in the biomedical domain. While other languages exist, largely due to historical reasons, they almost always have a translation to OWL.

For example, the OBO Foundry has been active since before OWL had been fully specified, and as such they have developed an ontology language of their own, called OBO (while the name is the same, OBO Foundry and OBO language are distinct concepts). With the standardization of OWL, however, OBO language has been almost completely deprecated: its constructions can be represented in OWL (*i. e.* OBO is, semantically, a subset of OWL), and tools to convert from it to OWL have been developed.

Since my work was based on OWL ontologies, it is important to let the reader know some of the terminology and notation used by this descriptive language.

In general, OWL ontologies are a representation of the concepts that describe the domain of knowledge being encoded in the ontology. The basic object of an OWL ontology is therefore the *class*: the representation of the real-world concept. Another important notion is the *individual*, which is the representation of a real-world object relevant for the domain. For example, a human anatomy ontology would contain the class Heart, which can be thought of as the set of all the individual "human hearts": my heart, the reader's, and all other human hearts that have existed, exist in this moment, and will exist in the future.

OWL ontologies also make use of *properties*, which are the "verbs" that represent the relationship between the individuals. For example, "my heart" *has-part* "my left ventricle". Properties are exclusively asserted between individuals; however, the ontology can describe high-level collections of property assertions that are known to be true. For example, since all human hearts have one left ventricle, a human anatomy ontology contains an *axiom* that represents this fact (see **Figure 2.2**).

There are several types of axioms that can be asserted in an OWL ontology, which are the

constructions of the ontology most closely based on the field of description logic and deductive reasoning. The types of axiom relevant for this work are the following:

- The *subclass of* axiom states that all the instances of one class are also instances of the other class: *e. g.* "Ventricle $\sqsubseteq$ Cardiac chamber" means that all ventricles are cardiac chambers.
- The *disjointness* axiom between two states that there can never be an object that is simultaneously an instance of the two: *e. g.* "Ventricle $\sqcap$ Atrium $\sqsubseteq \bot$" means that there is no object in the real world that is both a ventricle and an atrium.
- The *existential quantification* axiom states that instances of a class are related to an instance of another class by a certain property: *e. g.* "Heart $\sqsubseteq \exists$ *has-part* . Left ventricle" means that every heart has a part that is a left ventricle.

The logical nature of these axioms are the main reason that ontologies such as MeSH are not on the same formality level than other carefully crafted logical ontologies. For example, a hypothetical axiom "Head $\sqsubseteq \exists$ *has-part* . Nose" (all heads have a nose), together with the fact that Ear is categorised under Head in this ontology, would lead to the incorrect inference that every ear has one nose.

While OWL allows the description of individuals, biomedical ontologies are developed and used as reference ontologies for various purposes and, as such, do not define any particular instance of their classes. They only describe the knowledge at the high level of the concepts.

Ontologies in general, and OWL ontologies in particular, make use of what is known as the *open-world assumption*. Informally, this assumption states that what is not asserted does not give any information about what is known *not* to be true. One consequence is that if an ontology does not contain subclasses for a given concept, it cannot be assumed that no such subclasses exist. A highly appropriate quote from Martin Rees [OB71; Ber73] perfectly encompasses this assumption:

> "Absence of evidence is not evidence of absence."

This has significant impact on the rules of inference that are allowed in OWL ontologies, and has consequences to the overall research performed in this area, as we will see later in Chapter 5 "Towards OWL-aware similarity".

Being part of an effort to make knowledge more accessible to machines, OWL language uses the idea of *universal identifiers*. The Internationalized Resource Identifier (IRI) are, superficially, similar to URLs. For example, the most abstract concept that can be used in an OWL ontology is http://www.w3.org/2002/07/owl#Thing, a class that contains all instances. This identifier can be used in any OWL ontology with the meaning defined by the World Wide Web Consortium (W3C). Likewise, any identifier that is a valid IRI can be used by any OWL ontology, and the universal nature of the identifier assures both developer of the ontology and its users that the class represents the concept defined in the ontology where it was originally created.

2.4 Semantic web

Once knowledge has been made machine-readable by using ontology concepts and entity annotation, it needs to be stored and shared amongst interested parties. This idea of publishing and sharing machine-readable information has been made possible by the semantic web, which prescribes both *i*) a set of standard formats for representing knowledge (of which OWL and IRI are examples); and *ii*) a collection of technologies to deal with knowledge (such as reasoning over OWL ontologies). In particular, the semantic web is a vision of information management and sharing that promotes intelligent access to data on the World Wide Web, both by humans and by computers [BHL01; SBH06]. It is especially useful for handling heterogeneous data, since it was designed with a structured yet flexible operation mode.

The semantic web is build around the idea of expressing information in structured and formal languages, such as the Resource Description Framework (RDF), that allow the expression of precise statements (*e. g.* "Mary" *has-father* "Peter"). At the most basic layer, the semantic web does not define what the property *has-father* means, working instead as a framework for sharing formal statements, which allows users with the necessary knowledge to deal with this information according to their needs. At a higher layer, semantic web *does* use the expressive and logical power of ontologies, enabling data to be effectively searched based on its semantics rather than its syntax. For example, with an ontology that contains the fact that the property *has-father* is the inverse of *father-of*, a user can search in a data repository for the objects x that satisfy the expression "Peter" *father-of* x, and still find the answer "Mary" (along with all her siblings, if any exist in the repository): while this exact statement was never introduced in the repository, the inverse statement was, and the relation between the properties *has-father* and *father-of* allows the search engine to correctly *infer* this answer. This illustrates one of the most important characteristics enabled by the semantic web: interoperability of data. On the one hand, information is shared using standard formats; on the other hand, the semantics associated with the information, *i. e.* the meaning and the implications of the data, are formalised based on the precise semantics of the languages used to describe it. This enables data owners to describe their data using as much precision as deemed necessary, while allowing researchers to query the data being as general as they want, while still guaranteeing that the relevant information is retrieved.

The use of reasoners enables the production of new data, and allows computers to process the structured information based on their actual semantics. An example of semantic web in action can be seen in the work by Lopes and Oliveira [LO12]. These authors have developed a framework capable of integrating structured knowledge from various sources in a single platform, which is enriched with web services that enable *knowledge federation*, *i. e.* the possibility to query data wherever it resides, without the need to add it to a local repository. The formality behind semantic web data suggests that these data can be linked with other information [BHB09; Biz09], just like documents in the web are linked to each other.

```
PREFIX epidemic: <http://www.epidemiology.com/data/>
PREFIX geo: <http://www.geography.com/data/>

SELECT ?characteristic
WHERE {
    epidemic:H1N1_surge_of_2009 epidemic:surge_started_in ?location .
    ?location geo:has_characteristic ?characteristic .
}
```

Listing 2.1 – Finding the characteristics of the starting place of an epidemic with SPARQL. This query retrieves the information we are looking for. Notice that it depends on non-existing repositories (http://www.epidemiology.com/data and http://www.geography.com/data) and, as such, is not functional. Even if these repositories existed, the query would only work if `geo:has_characteristic` was a superproperty of all the relevant properties.

An interesting example of the semantic web in action is the use of linked data to cross information on some epidemiological surges with the characteristics of the locations where these surges started (*e. g.* the socio-economic or environmental conditions). To illustrate, consider a collection of epidemiological surges together with a repository containing characteristics of geographical locations. The search presented in **Listing 2.1** is written in SPARQL, a language that expresses queries over RDF stores (also a standard proposed and promoted under the semantic web movement [HS13]). If presented to the correct data repositories, it would return the characteristics of the places where the "H1N1 surge of 2009" started. Then, comparing the returned information with the results for other epidemic surges, it would be possible to detect the characteristics more strongly associated with each one, and to find patterns in the data.

Another relevant example of semantic web in action is the Open PHACTS project [Wil12], which provides an integrated and interoperable platform that aims at reducing barriers in pharmacology, specially in the task of drug discovery. The general methodology followed for this endeavour is the adoption of semantic web technologies, such as RDF stores, semantic annotation (see next section), SPARQL queries, *etc.* which are integrated in the platform, thus building on open standards to ensure wide applicability of the approaches used for integration of data.

2.5 Semantic annotation

Ontologies, standing on their own, define a set of unambiguous, objective and traceable concepts, along with their names, synonyms, and (formal or textual) definitions. However, as knowledge artefacts, ontologies do not *do* anything. Using an analogy, ontologies are to knowledge as the source code of a program is to the program itself. They are specifications packed with a lot of potential, and liberating this potential is possible only with the right set of tools, giving researchers the ability to explore the knowledge they contain. Therefore, it is essential for the advancement of science that the community develops and uses ontologies in a way that can be stacked with current technologies designed for this area. In fact, the knowledge that is stored in

an ontology can be quite expressive, depending on its format and how formal its representation is, and can be explored in various ways.

For example, repositories enriched with ontology axioms can be paired with SPARQL to allow intelligent search of data within the repository (see **Listing 2.1**). We will see in Section 2.6 "Semantic similarity" that another such technology is semantic similarity (the primary subject of this document), which calculates similarity between entities based on the knowledge that is associated with them.

Right before discussing semantic similarity, however, it is important to understand *what* is usually compared with this technique. Comparing concepts with concepts is not always useful, and ideally we would like to compare full entities (clinical notes, proteins, disease, *etc.*) which are usually *annotated* with ontology concepts but are not themselves concepts. For example, a common practice in biology is annotating proteins with their functions. It is almost universally accepted that protein functions are well represented in the Gene Ontology (GO). With this ontology, the information that a gene is responsible for a specific function can be expressed, *e. g.* the protein "telomerase" UniProt:Q99973 is annotated in the UniProt database as having the function "ATP binding" GO:0005524 and being localised in the "nuclear matrix" GO:0016363). This statement is objective, unambiguous (*i. e.* it does not depend on the researcher that made the statement, nor on any other context), universal, and traceable. Databases like AmiGO [Car09] are dedicated to managing statements like this.

These annotations can be seen as a semantic description of the protein, since they can be used to, computationally, ascribe to the protein a meaning more complex and informative than simply its sequence. There are automatic tools that reason over GO annotation in order to help interpret the results of experimental procedures. For example, Gene-Set Enrichment Analysis determines, based on the gene expression levels in a wild-type individual *vs.* those of a mutated individual, which GO molecular functions are most strongly associated with the mutated individuals [Sub05]. This can help identify, for instance, molecular causes of a disease. Relying on ontology concepts to annotate biomedical entities allows automatic reasoning to be applied directly to them, increasing the amount of automation that can in theory be applied in biomedical research.

In many cases, the annotations of an entity span more than one ontology. In epidemiology, a single dataset may require annotation with diseases, geographical locations, medical procedures, socio-economic conditions, *etc.*; kinetic models of chemical reactions use concepts representing chemical compounds and the mathematical equations for the reaction's velocity. Given this multidisciplinarity, it is essential that the standard ontologies used throughout the community can work together, providing users with the confidence that their annotations are interoperable. As discussed previously, this is the case with most ontologies of the biomedical domain. For example, some proteins capture ethanol molecules, a function represented in GO with the concept Ethanol binding. This concept is related to Ethanol, itself represented in CHEBI. Such interoperability also has the advantage of minimising the risks of representation duplication (akin to "code duplication" in software development).

Finally, semantic annotation is itself a form of knowledge representation. By stating, in a machine-readable format, that some protein performs a certain function in the cell, we are augmenting the amount of knowledge that can be exploited by computational methods.

2.6 Semantic similarity

Now that some preliminary concepts have been introduced, we are finally ready to appreciate the notion of semantic similarity.

Traditionally, computers have been able to compare objects that can be represented either mathematically (*e. g.* vectors) or as strings of characters (*e. g.* gene sequences). However, the algorithms that are used with these structures are context-free: they usually transform the structures without any knowledge of what they represent. With the help of a formal representation of knowledge, computers are given the ability to manipulate concepts that are difficult to represent in a mathematical way.

Knowledge representation (by means of ontologies and semantic annotation) provides the appropriate support for automatic manipulation of information. In this context, semantic similarity is a technique that assigns a numeric value to a pair of concepts or annotated entities based on the similarity of their *meaning*, which is automatically extracted from the ontologies.

For example, there is no directly obvious way to compare two anatomical entities. However, considering the illustration in **Figure 2.2** (page 17), it is possible to intuitively understand that, because both a Heart and a Stomach are examples of a Cavitated organ, they are more similar than Heart and Pancreas. This intuition can be captured in a formal algorithm: Heart and Stomach are both subclasses of the concept Cavitated organ, while Heart and Pancreas are subclasses of the concept Organ, a less specific concept. The fact that this measure of similarity makes use of the meaning of the concepts, as represented in the ontologies, has impelled the use of the phrase "semantic similarity", first used in this context by Resnik [Res95]. Although the meaning of a concept is also non-mathematical, it is possible to use ontologies as proxy for that meaning and KR technologies to manipulate it. For this reason, semantic similarity can also be called "ontology-based similarity". For the purpose of this thesis, I define semantic similarity as follows:

> "A semantic similarity measure is an algorithm that takes as input a pair of ontology concepts (*resp.* a pair of entities annotated with ontology concepts), and returns a numeric value that reflects how similar the concepts (*resp.* entities) are; the meaning of the concepts being compared (*resp.* used to annotate the entities) is retrieved from the ontologies where they are defined."

Semantic similarity has been applied in several areas of research. Hoehndorf et al. [HDG13] provide a collection of applications that contribute to verifiable scientific advances. Some examples collected by me during my research are:

- predicting protein interactions (either physical interaction, as part of the same complex, or less obvious interactions, like being part of the same metabolic pathway) [AB04; Guo06; Wu06];
- predicting sub-cellular location of proteins [LD06];
- predicting whether a disease affects a certain body part [FC11];
- finding protein complexes in protein-protein interaction networks [Li10b];
- helping the differential diagnosis process by suggesting diseases based on a set of symptoms [Köh09];
- predicting chemical properties in small metabolites [FC10];
- finding new putative uses for drugs that are currently being used (drug repositioning) [Tan14];
- assisting visualization techniques by finding representative concepts in a large set [Sup11];
- being part of large information retrieval systems [Ema14];
- determining the meaning of ambiguous terms [McI11; Hu12; GCA14];
- improving the classification of clinical texts based on machine-learning [GB12]; and
- assisting text-mining by providing a means to detect similarities in meaning that are not obvious using string-similarity measures [Spa05; Var05] and by disregarding some mined facts if they fail to verify a constraint on semantic similarity [LFC15].

The notion of similarity it tightly coupled with the notion of *relatedness*. From a technical point of view, similarity and relatedness are the same idea: they assign a value to a pair of concepts/pair of annotated entities. As such, distinguishing between the two ideas is generally difficult. As a rule of thumb, it has been proposed that similarity is context-independent (it takes into account the concepts being analysed but disregards the application under which they are being compared) and relatedness depends on the goals behind the analysis [Bud99]. Pedersen et al. [Ped07] were amongst the first to make a more formal distinction between these two ideas: similarity is a special case of relatedness that considers only the hypernymy of concepts (the class-subclass hierarchy), while relatedness explores all other kinds of properties in the ontologies.

Take, for instance, the concepts Heart and Blood. In the biomedical field, they are closely related, since the function of the former is to pump the latter. In gastronomy, there is little relatedness between the two. Independently of the context, however, a heart is not at all similar to blood: one is an organ, the other a biological fluid (in fact, a liquid tissue).

In fact, in some contexts, the use of properties other than the class-subclass relationship can be useful to find patterns in data or to infer new conclusions. Consider the graph drawn in **Figure 2.3**. This figure was obtained by drawing an edge between two anatomical concepts from FMA (the Foundational Model of Anatomy) if the two are related by means of the property *articulates-with*: hence, each node is a bone and the edges show that there is an articulation

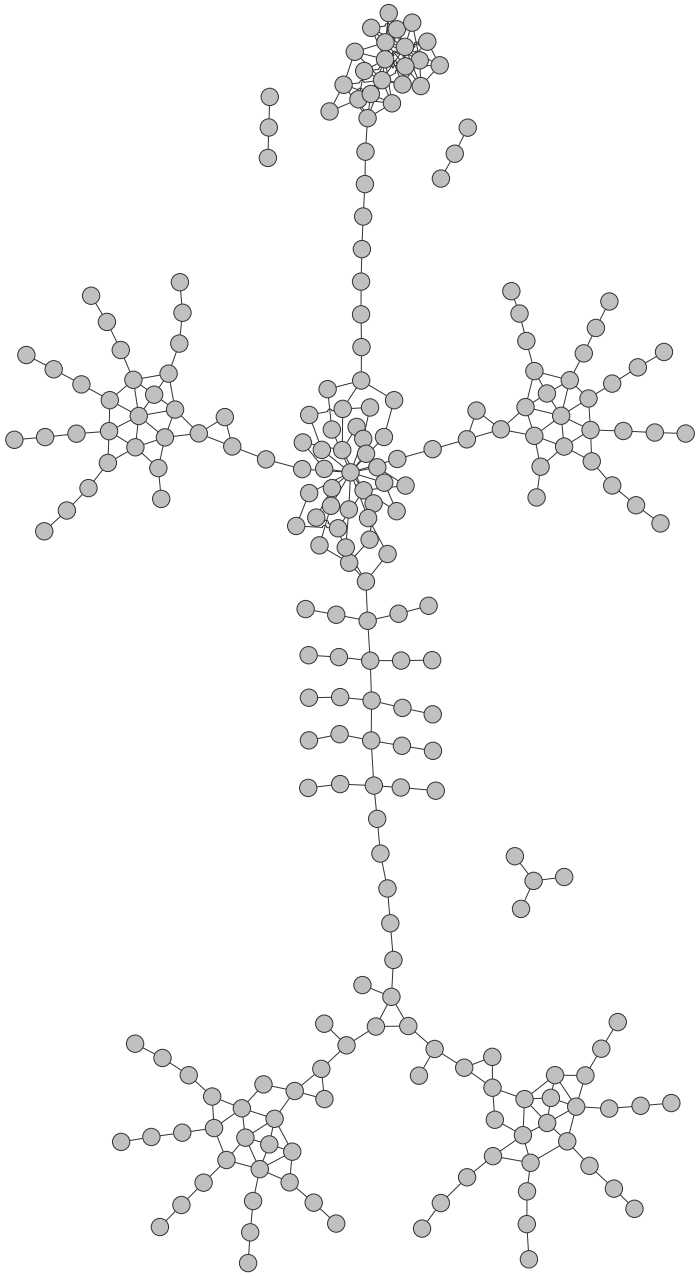


Figure 2.3 – Graph describing the *articulates-with* property in FMA. Each node is a concept from FMA representing a human bone, and each edge represents the fact that the two concepts are related to one another by means of the *articulates-with* property. Features of the human body visible in this picture include: the head, the rib cage, the two hands and their fingers, the "false ribs" (which do not articulate with the sternum), the vertebrae, the two feet and their toes, and also the two sets of ear bones, which are disconnected from the rest of the body.

between the two bones. A graph editor was then used to automatically layout the nodes (I used yEd). Finally, minor manual adjustments were carried out, both to rotate the image and to move the three disconnected subgraphs. Upon closer inspection, it is evident that the graph mirrors a slightly deformed human body, with a head, two hands, a spine, and two feet: the head is tangled because most bones articulate with a large set of other bones, and the rib cage is also tangled because the upper ribs connect both to the spine and the sternum (the two disconnected pieces on the top are the bones of the ear; the third disconnected piece is a bug in the ontology). That this picture is obtained with minimal human intervention and that, even so, it so closely resembles the human body, which is the object being represented in FMA, is extremely provoking evidence that relationships other than hypernymy are vital to properly process scientific knowledge.

In the context of multi-domain similarity, relatedness has, in fact, high utility. Consider now the concepts Otitis and Ear, likely to be represented in different ontologies: one for diseases and the other for anatomical entities. Despite being from different ontologies, there is a strong relation between the two, as otitis is an infection of the ear. In the context of diseases, this relationship increases the relatedness between Otitis and other ear diseases (*e. g.* Hereditary deafness), which is difficult to capture using similarity alone, since an Otitis is an inflammatory disease and Hereditary deafness is a genetic disease. In such a disease hierarchy, it is impossible to obtain an accurate comparison value between the two diseases based on the class-subclass hierarchy only. In fact, we know that Otitis is an "Inflammatory disease that *is-located-in* some Ear" and that Hereditary deafness is a "Genetic disease that *manifests-in* some Ear". Therefore, exploring the relationships between the concepts other than the class-subclass hierarchy can help increase the accuracy of semantic similarity.

The concept of distance measures should also be mentioned. Before ontologies were used to compute similarity measures, they were used to compute distances between concepts [Rad89]: for a pair of concepts that are close in meaning, similarity values are high and distance values are low. Although there is no unique way to convert a distance into similarity, some formulae have been frequently used. Denoting distance with d and similarity with σ, they are

- $\sigma = 1/d$;
- $\sigma = D - d$ where D is the maximum possible distance, and
- $\sigma = e^{-\gamma \cdot d}$ for some $\gamma > 0$.

2.7 Multiple-ontology context

As with other emergent fields, the practice of ontology development has been tackled by various people, from hobbyists to philosophers, from scientific research teams that need the power of ontologies for their research, to enterprises that sell their knowledge representation of reality. This can lead to many different ontologies being constructed by different people, with either a different philosophical base or simply a different perspective of reality.

The dissemination of ontology construction and usage places the utility of ontologies in

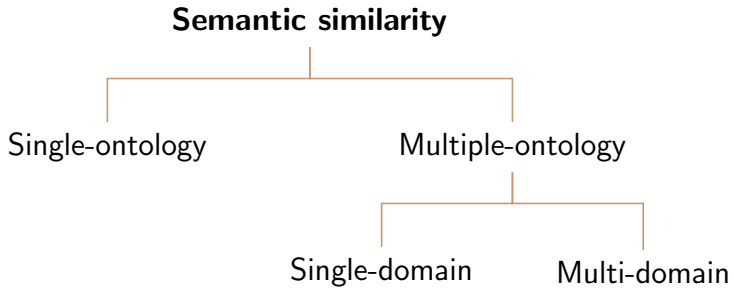


Figure 2.4 – Categories of semantic similarity. Semantic similarity (and relatedness) can be made in a single- or multiple-*ontology* context; but using more than one ontology does not immediately imply a multi-*domain* context, since some ontologies can model the same domains of reality (for example MeSH and NCIt overlap in some of their concepts). As such, multiple-ontology analysis can be further subdivided into single- and multi-domain.

a vantage point, mainly due to two facts: *i*) different interpretations of reality can lead to complementary ontologies, and *ii*) a variety of domains of knowledge are getting represented as ontologies, especially by those more competent to do so, *i. e.* people with a background knowledge in these domain.

With such a plethora of ontologies available, it is not surprising to notice that many applications are now making use of more than a single ontology. For example, the Epidemic Marketplace [Lop10] uses a number of epidemiology-related ontologies to annotate its resources [Fer12], thus connecting a web resource to various concepts from different domains. Another example are models of biological systems, also being annotated with multiple ontologies: with the processes they model, the chemical molecules and cellular components involved in those processes, the physical quantities that they model, *etc.* [Li10a; Jut15]. To properly achieve a significant semantic similarity measure between such multidisciplinary entities, it is essential that a multi-domain measure be developed.

Consider the categorization of semantic similarity methods illustrated in **Figure 2.4**. Both concepts Heart and Blood can be included in an ontology of anatomy and, therefore, their similarity can be computed with a single ontology. However, biomedical ontologies are often incomplete, due to the intrinsic uncertainty associated with the scientific field; they can also contain errors, or can even follow a certain view of reality that is not shared amongst everyone. In each of these cases, a second ontology may be used to offer a complementary view of reality so that incompleteness, errors and subjective interpretations are mitigated. "Multiple-ontology single-domain" similarity can be used to handle this situation [*e. g.* MN09]. In this approach, two or more ontologies representing the same domain are used in a complementary way to improve semantic similarity results.

Multiple-domain similarity represents a step beyond this approach, since it uses multiple ontologies from distinct domains in order to compare concepts in a multidisciplinary context. This is necessary, for example, when performing relatedness analysis, such as when comparing

diseases with symptoms, or symptoms with anatomical entities, but also when comparing entities that are annotated with concepts from other domains. For example, the concepts Heart and Blood, previously used in the example of single-ontology similarity, can be compared based on their functions, the symptoms they exhibit, or even, if appropriate, their use in gastronomy.

2.8 Summary

In this chapter, I exposed some of the most important concepts necessary to understand the rest of this document. I started by visiting the notions of knowledge representation and ontologies as computational artefacts, thus laying down a theoretical framework that enables the representation of unstructured information in a way that can be parsed and acted upon by machines. I then mentioned the ideas of semantic web and semantic annotation, which are the response of the scientific community to that theoretical framework: they define the standards that are used to store and share information among the ones representing the knowledge and the ones using that knowledge. At last, I described semantic similarity as one of the techniques that uses information from ontologies, with several possible objectives, and the fact that multi-domain semantic similarity seems to be useful and, yet, underdeveloped.

The next chapter will describe some of the methodologies that have been proposed to calculate semantic similarity, including both historical and state-of-the-art measures.

Chapter 3

State of the art

This section is an exposition of both an historical and current state-of-the-art in semantic similarity calculation. It starts by describing the first few measures and how they evolved through time. Throughout the chapter, I present both the concepts behind the measures of similarity proposed by various researchers and some selected formulas used by these measures.

3.1 The art of semantic similarity

The study of semantic similarity has been subject of research for a significant amount of time. A first work by Tversky [Tve77], published in *Psychological Review* laid the first steps in the formalism of the mathematical calculation of similarity by developing a theory that tries to explain similarity as judged by people. In this work, similarity is described as a function of the features of the things being compared, namely common features *vs.* distinctive features, *e. g.* shape for geometric figures or political aspects for countries.

A previous idea was published some years before by Quillian [Qui68] and Collins and Loftus [CL75] (these have no notion of similarity being calculated by computers, but instead lay down some theoretical psychological views on how people perceive similarity), which proposes that the mental processes by which humans organise their memories and concepts are based on a network of connected concepts, whose connections are stronger for more related concepts.

With the advent of computerised science, the idea that automatic systems could be able to compare concepts and other knowledge artefacts started to emerge, and thus the idea of semantic similarity was introduced.

3.2 Edge-based approaches

The works mentioned in the previous section have prompted Rada et al. [Rad89] to create a first measure of semantic distance based on a hierarchy of concepts, in this case the Medical Subject Headings (MeSH). They calculated distance as a function of the number of class-subclass relationships that must be traversed in order to go from one concept to the other in the hierarchy (also called the "edge distance"). For example, using the ontology snippet in **Figure 3.1**, the

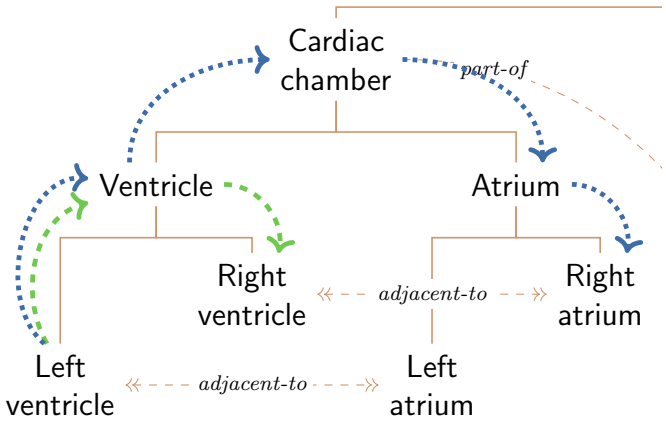


Figure 3.1 – Semantic measures explained in a hypothetical hierarchy. This small detail of the hypothetical ontology presented in **Figure 2.2** (page 17) includes further information representing one of the core ideas behind semantic measures. The concepts Left ventricle and Right ventricle are direct subclasses of Ventricle. As such, the edge distance between the two concepts is 2, as one has to climb up through the Left ventricle → Ventricle relationship and then down through the Ventricle → Right ventricle relationship to go from one concept to the other (see the green dashed arrows). Similarly, the edge distance between Left ventricle and Right atrium is 4 (blue dotted arrows).

edge distance between Left ventricle and Right ventricle is 2, and the edge distance between Left ventricle and Right atrium is 4. This vision of semantic analysis draws from the idea that an ontology can be represented as a tree, explained previously in Section 2.2 "Ontologies". In fact, the use of trees to represent ontologies has become so widespread in this area that it is customary to use the notions of "ancestors" (*resp.* "descendants") of a concept as the set of its direct and indirect hypernyms (*resp.* hyponyms), thus making Ventricle an ancestor of Left ventricle and a descendant of Cardiac chamber.

Edge distance can be converted into similarity as detailed in Section 2.6 "Semantic similarity". For example, the work by Pedersen et al. [Ped07] refers to the use of $\text{sim}(a, b) = 1/p(a, b)$, where $p(a, b)$ is the number of nodes in the shortest path between concepts a and b. Wu and Palmer [WP94] propose a normalization, using for that effect the notions of the *least common subsumer* (LCS) between a and b, which is the most specific concept that subsumes both a and b (*e. g.* the LCS between Left ventricle and Left atrium is Cardiac chamber) and *depth* of a concept, which is the edge distance between the root of the ontology and that concept:

$$\text{sim}_{\text{Wu}}(a, b) = \frac{2\delta(a, b)}{p(a, b) - 1 + 2\delta(a, b)} \tag{3.1}$$

where $\delta(a, b)$ is the depth of the LCS between a and b.

While intuitive, these approaches assume some things that are not always true in the biomedical domain: *i*) the amount of concepts and class-subclass relationships is uniform throughout the

various sub-domains represented in the ontology; and *ii*) class-subclass relationships at the same depth in the ontology correspond to the same semantic distance between the two concepts. In fact, concepts are denser where they represent more "appealing" areas of research, as that is where research focusses and, as such, where the community's knowledge is more detailed. A simple example shows these faults: the intuitive distance covered in "Fungal spore *is-a* Spore" seems to be narrower than the distance in "Plankton *is-a* Organism form" (examples taken from MeSH). Strategies have been proposed to attenuate these issues, such as weighting edges differently according to their hierarchical depth, or to node density and type of link [Pes09b].

3.3 Node-based approaches

Because of the problems mentioned above, focus changed from the edges to the nodes of the graph representation of the ontology. Resnik [Res95] proposed an information-theoretic notion called Information Content (IC), which depends on the frequency with which this concept is used to annotate entities in a corpus. For example, Organ can refer to many distinct concepts, and, as such, carries a small amount of information when compared to the concept Heart, which has a more informative definition; this measure, thus, reflects the *specificity* of a concept. Resnik [Res95] has shown that measures that use the notion of IC to weight the concepts of an ontology perform better than those that rely on edges alone. (Note, however, that IC-based measures are also biased, as the annotation process is guided by the trends in research and, as such, there are more annotations made to concepts more related to "hot" research topics.)

The idea of using IC in semantic similarity is that it allows the definition of *shared information* between concepts. Reusing the example from **Figure 3.1**, Left ventricle and Right ventricle share between themselves the definition of Ventricle (*i. e.* Ventricle is a common superclass of both). Since IC reflects specificity, the similarity between two concepts can be computed as the IC of their most informative common superclass. This results in the intuitive notion that Left ventricle is more similar to Right ventricle (both are Ventricles) than to Right atrium (they share only the fact that they are both Cardiac chambers), since Ventricle is more specific than Cardiac chamber.

The original work by Resnik [Res95; Res99] defined the information content of a concept c based on ideas from information theory:

$$\mathrm{IC}_{\mathrm{Resnik}}(c) = -\log f(c) \tag{3.2}$$

where $f(c)$ is the frequency with which concept c appears in a selected corpus. For example, in WordNet, a taxonomy of English words [Mil95], the number of occurrences of a concept is counted as its frequency in a collection of texts; for GO, Lord et al. [Lor03] measured the frequency of a concept as the number of proteins in the SwissProt database that are annotated with that concept. It is important to notice that a reference to a particular concept, *e. g.* Left ventricle, is also a reference to its hypernyms, in this case Ventricle, Cardiac chamber and Anatomical structure (*cf.* **Figure 2.2** on page 17). With this rule, it becomes trivial to prove that as one

moves from abstract to specific concepts, the IC increases, as is expected for any measure of specificity.

The notion of information content, however, need not necessarily require the exploration of external corpora. It is possible to measure the specificity of a concept based on the structure of the ontology itself: for example, the number of concepts that are subsumed by c is intuitively higher for less specific concepts, while the concepts with no subclasses (sometimes called the *leaves* of the ontology) are the most specific concepts. Seco et al. [SVH04] use this idea in order to define an *intrinsic* measure of information content:

$$\mathrm{IC}_{\mathrm{Seco}}(c) = 1 - \frac{\log N_d(c)}{\log N} \tag{3.3}$$

where $N_d(c)$ is the number of direct and indirect hyponyms of concept c (including c itself) and N is the total number of concepts in the ontology. As previously, more abstract concepts will have a lower IC value. This measure is adapted from eq. 3.2 by taking $f(c) = \frac{N_d(c)}{N}$ and normalising so that the highest possible IC is 1.

The main advantage of intrinsic IC measures is that they are independent of external resources, and, therefore, can be calculated using the ontology alone. A review by Sánchez et al. [SBI11] shows that intrinsic methods do, in fact, correlate better with human judgement of similarity. This evaluation, however, is done on a set of 30 pairs of concepts from WordNet. This is a very small number of pairs to use in an evaluation, given the size of WordNet; additionally, given its domain, WordNet is, in some senses, different to the ontologies used in biomedical research, as concepts in WordNet have a collection of meanings (many English words have, in fact, more than one definition), while the concepts from biomedical ontologies strive to be unambiguous. As such, these results may not be true for the biomedical domain.

Using such measures of specificity, it is possible to estimate the similarity between two concepts as the IC of their most informative common ancestor:

$$\mathrm{sim}_{\mathrm{Resnik}}(a, b) = \max_{c \in \mathrm{CA}(a,b)} \mathrm{IC}(c) \tag{3.4}$$

where $\mathrm{CA}(a, b)$ is the set of all hypernyms common to both a and b. This was in fact the first node-based measure ever proposed [Res95]. For example, in the ontology in **Figure 2.2** (page 17), $\mathrm{CA}(\mathsf{Heart}, \mathsf{Stomach}) = \{\mathsf{Cavitated\ organ}, \mathsf{Organ}, \mathsf{Anatomical\ structure}\}$ and since Cavitated organ has the highest IC in this set (it is the most specific), $\mathrm{sim}(\mathsf{Heart}, \mathsf{Stomach}) = \mathrm{IC}(\mathsf{Cavitated\ organ})$. The notion of *most informative common ancestor* (MICA) is so widespread that it has its own mathematical definition:

$$\mathrm{MICA}(a, b) = \operatorname*{arg\,max}_{c \in \mathrm{CA}(a,b)} \mathrm{IC}(c). \tag{3.5}$$

Likewise, the idea of measuring the shared information content between two concepts as the IC of their MICA is also so common that I use a notation for that as well:

$$\mathrm{IC}_{\mathrm{s}}(a, b) = \mathrm{IC}(\mathrm{MICA}(a, b)). \tag{3.6}$$

As happened previously with edge-based measures, the idea presented in eq. 3.4 has been subsequently adapted by other authors in order to solve some of the problems it presents: *i*) the measure is unbounded when it uses internally an unbounded IC measure (such as the one in eq. 3.2), and *ii*) the similarity between two specific concepts whose MICA is some concept c is the same as the similarity of two abstract concepts whose MICA is also c (*e. g.* the pairs Left ventricle/Left atrium and Ventricle/Atrium in **Figure 2.2** are equally similar using $\text{sim}_{\text{Resnik}}$, but this notion is contrary to general human intuition). Solving the first issue is a matter of normalising the measure of IC (*e. g.* dividing it by the maximum possible IC) [Pes08], but the second issue remains. Lin [Lin98] introduced a normalization approach that prevented both issues:

$$\text{sim}_{\text{Lin}}(a, b) = \frac{2 \times \text{IC}_\text{s}(a, b)}{\text{IC}(a) + \text{IC}(b)}. \tag{3.7}$$

Another approach, by Jiang and Conrath [JC97], defines a distance measure instead of a similarity one, using a normalised measure of IC:

$$\text{d}_{\text{Jiang}}(a, b) = \text{IC}(a) + \text{IC}(b) - 2\,\text{IC}_\text{s}(a, b) \tag{3.8}$$

which can be converted to similarity with $\text{sim}(a, b) = 1 - \text{d}_{\text{Jiang}}(a, b)/2$ [Li11; Bat12], or $\text{sim}(a, b) = 1/(\text{d}_{\text{Jiang}}(a, b) + 1)$ [CSC07]. See Section 2.6 "Semantic similarity" for more ways to convert distance into similarity.

Other node-based measures of similarity exist that do not take into account the information content of the concepts. For example, Sánchez et al. [Sán12b] defined a distance measure that takes into account the number of common ancestors between the two concepts. They use the function $\text{A}(c)$, which returns the set of hypernyms of concept c and define:

$$\text{d}_{\text{Sánchez}}(a, b) = \log_2 \left(2 - \frac{|\text{A}(a) \cap \text{A}(b)|}{|\text{A}(a) \cup \text{A}(b)|}\right). \tag{3.9}$$

More recently, there have been some works dedicated to the augmentation of the notion of information content, especially *shared* information content (IC_s). Couto and Silva [CS11] have developed the *disjunctive information content* (DiShIn), a measure of shared information content that takes into account the fact that concepts can have more than one parent. This measure can be considered a *plug-in* that relies on other IC measures and introduces the new capability by shaping the IC_s measure according to whether the two concepts being compared have multiple *disjunctive* ancestors (the definition of which is beyond the scope of this document, but it is based on the set of common superclasses that are not superclass of each other). In this case, the shared IC between concepts a and b is the average of the IC of all disjunctive ancestors:

$$\text{IC}_\text{s}(a, b) = \frac{1}{|\text{DCA}(a, b)|} \cdot \sum_{i \in \text{DCA}(a,b)} \text{IC}(i) \tag{3.10}$$

where $\text{DCA}(a, b)$ is defined as the set of all disjunctive common ancestors between the two concepts.

3.4 Semantic relatedness

As previously stated, similarity and relatedness are two different ideas (Section 2.6 "Semantic similarity"). Theoretically, similarity is a specific case of relatedness that uses the hypernymy relationship, while relatedness uses all the properties between the concepts [Ped07].

In practice, however, little distinction has been made in the literature between these two notions. For instance, most measures of semantic similarity applied to GO assume that the whole-part relationship is equivalent to class-subclass relationship. In the first work to apply semantic similarity in GO [Lor03], the authors assume that the ancestors of Nucleus include the concepts Cell, even though the nucleus is part of the cell, not a subtype thereof. Likewise, semantic similarity in CHEBI [Gre10; FC10] uses properties like *has-functional-parent* and *has-role* to define the ancestry of a concept, but never acknowledge the difference between similarity and relatedness.

In the biomedical domain, there seems to be a lack of research in the area of relatedness. Pedersen et al. [Ped07] present a measure of relatedness between medical concepts from the SNOMED-CT ontology but, this measure is not ontology-based, since it calculates relatedness between two concepts based on the words that are frequently found, in text, around the two concepts being compared, and then comparing these two sets of words.

My generic relatedness measure validated in an ontology of human anatomy is among the first truly ontology-based semantic relatedness measures in the biomedical field (see Section 5.2 "Semantic relatedness measure").

3.5 Comparing annotated entities

Semantic similarity is not only about comparing concepts. As per the definition in Section 2.6 "Semantic similarity", it also handles comparison of annotated entities.

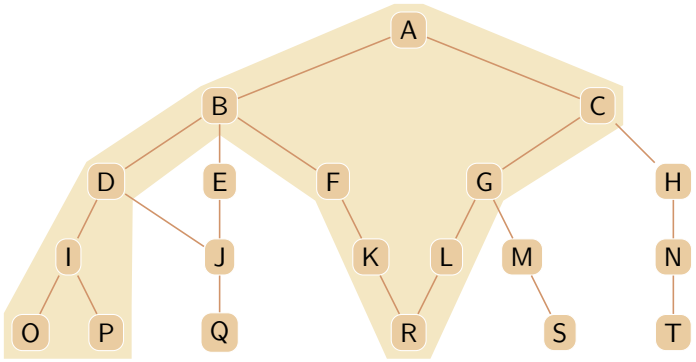
Some group-wise measures have been developed that can compute the similarity between two sets of concepts. For example, Gentleman [Gen07] compares entities using their annotations by first defining an *induced graph* for a entity, which is the graph containing the concepts that annotate the entity plus all their ancestors, up to the root of the ontology. Let $\phi(e)$ be the set of all concepts in the induced graph of the entity e; the formula proposed by this author to compare entities e and e' is:

$$\mathrm{sim}_{\mathrm{UI}}(e, e') = \frac{|\phi(e) \cap \phi(e')|}{|\phi(e) \cup \phi(e')|}. \tag{3.11}$$

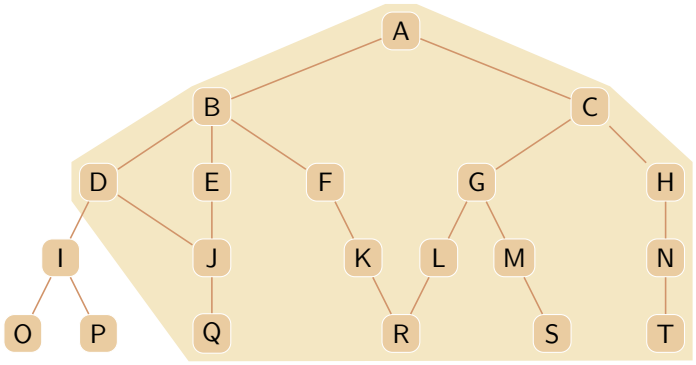
Pesquita et al. [Pes08] use a related approach to define their semantic similarity measure, called $\mathrm{sim}_{\mathrm{GIC}}$, using a related formula that weights each concept by their IC:

$$\mathrm{sim}_{\mathrm{GIC}}(e, e') = \frac{\sum_{i \in \phi(e) \cap \phi(e')} \mathrm{IC}(i)}{\sum_{i \in \phi(e) \cup \phi(e')} \mathrm{IC}(i)}. \tag{3.12}$$

Consider the toy ontology in **Figure 3.2** and two entities: e is annotated with the set $\{\mathsf{O}, \mathsf{P}, \mathsf{R}\}$ and e' with the set $\{\mathsf{Q}, \mathsf{R}, \mathsf{S}, \mathsf{T}\}$. The induced graphs of e and e' are shown in



(a) The induced graph for an entity annotated with concepts O, P and R.



(b) The induced graph for an entity annotated with concepts Q, R, S and T.

Figure 3.2 – Group-wise measures in action. Two entities, annotated with a set of concepts form a toy ontology, are being compared using the idea of the induced graph. The figures show the ontology concepts and the class-subclass hierarchy, as well as the induced graph for the two entities in a shaded background.

the shaded regions of the graphs in **Figure 3.2(a)** and **Figure 3.2(b)**, respectively. In this example, we have:

$$\begin{aligned} \phi(e) \cap \phi(e') &= \{\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}, \mathsf{F}, \mathsf{G}, \mathsf{K}, \mathsf{L}, \mathsf{R}\} \\ \phi(e) \cup \phi(e') &= \{\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}, \mathsf{E}, \mathsf{F}, \mathsf{G}, \mathsf{H}, \mathsf{I}, \mathsf{J}, \mathsf{K}, \mathsf{L}, \mathsf{M}, \mathsf{N}, \mathsf{O}, \mathsf{P}, \mathsf{Q}, \mathsf{R}, \mathsf{S}, \mathsf{T}\} \end{aligned}$$

Using Seco's intrinsic measure of information content from eq. 3.3, we get, for example, $\mathrm{IC}(\mathsf{B}) = 1 - \frac{\log 11}{\log 20} = 0.20$; and through eqs. 3.11 and 3.12, we can compute the similarity between the two entities:

$$\begin{aligned} \mathrm{sim}_{\mathrm{UI}}(e, e') &= 0.45, \\ \mathrm{sim}_{\mathrm{GIC}}(e, e') &= 0.34. \end{aligned}$$

Besides this type of measure, where the sets of annotations are directly compared to each other, there are group-wise measures that are built over concept-wise measures. For this purpose,

let $e_1, e_2, \ldots$ be the concepts that annotate a certain entity e. To compare two entities e and e' in this case, all the pairs (e_i, e'_j) are compared using the concept-wise measure, creating a similarity matrix which is then converted into a single value. For example, the average over all the similarities and the maximum of these values are used in the literature [Lor03; Gre10].

Another way of aggregating the matrix is the Best Match Average (BMA). This method goes through each of the e_i concepts and finds the annotation in e' that is most similar to it, and then goes through each of the e'_j concepts and finds the annotation in e that is most similar to it. The result is the average of all these maximum values [Wan07; Pes08; Gre10]. Assuming e has n annotations and e' has m annotations:

$$\mathrm{sim}_{\mathrm{BMA}}(e, e') = \frac{\sum_{i=1}^{n} \max_j \sigma(e_i, e'_j) + \sum_{j=1}^{m} \max_i \sigma(e_i, e'_j)}{n + m} \tag{3.13}$$

where σ is a concept-wise similarity measure (it takes as input two concepts).

Figure 3.3 illustrates this process using the same two entities that were used in the previous example. The similarity matrix, calculated with $\mathrm{sim}_{\mathrm{Resnik}}$ (eq. 3.4) and $\mathrm{IC}_{\mathrm{Seco}}$ (eq. 3.3), is shown in the depicted matrix, and the maximum values for each row and column are extracted (see the arrows in the picture) and averaged:

$$\mathrm{sim}_{\mathrm{BMA}}(e, e') = 0.42.$$

Lehmann and Turhan [LT12] suggest a further enhancement to this formula. One of the issues of BMA is that it only takes into account one annotation in e' for each annotation in e, specifically *the most* similar one. However, it can be the case that, for some i, e_i is extremely similar to two or more e'_j. Instead of extracting just the maximum similarity value in each row or column, they propose the use of a T-conorm [KMP04] to aggregate all the values. T-conorms are binary operations (commonly represented as $x \oplus y$) that satisfy a set of properties, the most important of which, in this context, being $\max(x, y) \leq x \oplus y \leq 1$. As such, it is possible to take into account more than just one most similar concept and to use the *trend* of concept similarity in order to calculate the similarity between the annotated entities.

For example, we can use $x \oplus y = x + y - xy$. Taking into consideration not just one value in each row/column of the matrix, but rather two values, we reach the situation described in **Figure 3.4**, where each row (*resp.* column) is aggregated to reach a final score that depends on the two highest values in that row (*resp.* column). In this case, we would have

$$\mathrm{sim}_{\mathrm{BMA}'}(e, e') = 0.51,$$

which is slightly larger than the previous calculation in eq. 3.14, because we are now taking into consideration more similarity values. This strategy produces higher values, where the increase depends on the overall trend followed by the similarity values of the matrix. For example, a matrix composed entirely of 0s and 1s will produce an identical value using the BMA approach or this one.

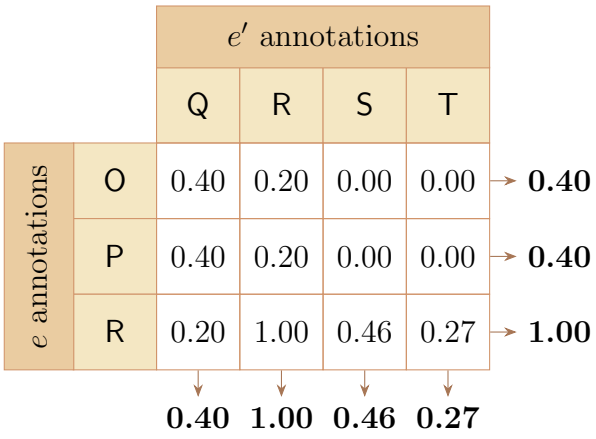


Figure 3.3 – Example of the best match average in action. Using the same entities referred to in **Figure 3.2**, this picture represents the best match average approach. Values in the matrix are the semantic similarity between the corresponding concepts, calculated with $\text{sim}_{\text{Resnik}}$ and IC_{Seco}. The arrows pointing to bold values show the maximum of each row and column, and the final result is the average of those maximum values.

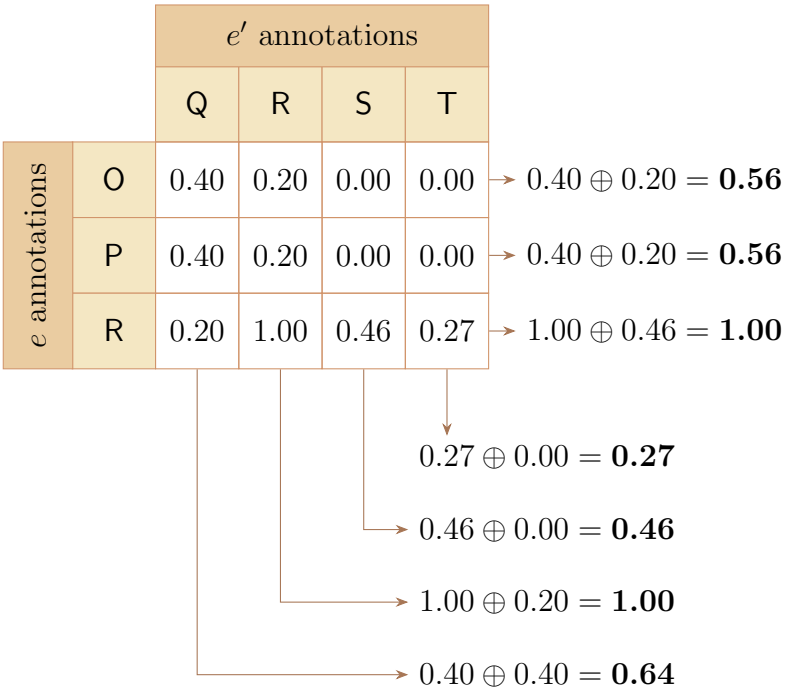


Figure 3.4 – The T-conorm aggregation strategy. This figure illustrates the use of the T-conorm $x \oplus y = x + y - xy$. This strategy uses the two highest values from every row and every column to derive a single value for the annotations in each of the two entities, rather than using only the highest.

3.6 Multiple-ontology semantic similarity

At the time of this document, and to the best of my knowledge, there is no published work dealing with semantic similarity in multi-domain contexts. Measures used in the literature are either single-ontology or multiple-ontology single-domain (see Section 2.7 "Multiple-ontology context") that use ontologies representing distinct perspectives of a single domain in order to enrich the similarity between the concepts of that domain. These studies are usually applied to the Systematized Nomenclature of Medicine "Clinical Terms"(SNOMED-CT) [Côt80], Medical Subject Headings (MeSH) [Rog63] and/or WordNet [Mil95].

While the absence of approaches in this area means that my own measures of multi-domain similarity will have nothing to be compared with, the multi-ontology single-domain approaches provide interesting insight into the issues of integrating multiple ontologies in a single measure of similarity and how to solve them.

These measures of similarity have been first approached by Rodríguez and Egenhofer [RE03]. Similarity measures that cross different ontologies rely on links between the concepts of the ontologies, meaning that it is vital that such links exist. The authors use lexical similarity to find these matches. While this is a weak form of ontology matching (for example, the lexical similarity between Role and Activity is 0 by most — if not all — intuitive lexical similarity measures, despite the fact that they are related concepts in chemistry), the use of synonymy enriches this measure. Semantic similarity is then calculated based on the class-subclass hierarchy of the concepts in the set of used ontologies, taking the links found in the previous step into account.

Al-Mubaid and Nguyen [MN09] propose an edge-based semantic similarity that also crosses ontologies. To do so, they introduce the notion of *primary ontology*: all similarities are calculated based on a scale that is balanced for the primary ontology. They use an edge-based measure and scale the edge distances of secondary ontologies according to the difference in maximum depth between the primary and secondary ontologies. The specificity of a concept, measured by its depth, is also appropriately scaled. Their similarity is then defined as

$$\mathrm{sim}_{\mathrm{AlMubaid}} = \log\Big((p(a,b)-1)^{\alpha}\cdot c_s(a,b)^{\beta}+k\Big) \tag{3.14}$$

where $p(a,b)$ is the scaled length of path between concepts a and b (which may cross ontologies by means of the inter-ontology links), $c_s(a,b)$ is the scaled depth of the LCS between a and b, and α, β and k are tunable parameters of the measure.

Sánchez et al. [Sán12a] define an approach that uses the notion of *semantic overlap*, which measures the overlap in the hyponyms of each concept:

$$\mathrm{overlap}(a^P,b^Q) = \frac{|H^P(a^P)\cap H^Q(b^Q)|}{\sqrt{|H^P(a^P)|\times|H^Q(b^Q)|}} \tag{3.15}$$

where c^O is used to represent the concept c in ontology O and $H^O(c^O)$ is the set of hyponyms of the concept c^O in ontology O. This overlap is calculated through lexical methods, as in the work of Rodríguez and Egenhofer [RE03].

Sánchez and Batet [SB13] extend the notion of information content to the multiple-ontology context, by defining the MICA of two concepts when *i*) the two simultaneously belong to two or more ontologies; or *ii*) there is no ontology that contains both concepts. This redefinition of the MICA is then used in the original formulas by Resnik [Res95] (eq. 3.4), Lin [Lin98] (eq. 3.7), *etc.*

As is evident from these works in multiple-ontology single-domain semantic similarity, using multiple ontologies to calculate similarity is a process that strongly depends on links between the ontologies. These can come from previous alignments produced by ontology matching techniques [ES07] or can be calculated *on-the-fly* by using lexical algorithms coupled with sets of synonyms. The recent advances in ontology matching, especially in the biomedical field [*e. g.* CAS09; Gro12], represent a big step forward in the determination of equivalent concepts between ontologies.

Given the ontology interoperability that is highly sought in the biomedical informatics community (see the last paragraph in Section 2.2 "Ontologies"), it is expected that concepts from two different ontologies never represent the same real-world idea. As such, ontology matching does not seem to be as useful in the context of multi-domain similarity as it is in single-domain similarity.

In contrast, what *is* necessary and useful is the notion of inter-domain links between concepts that are related. Some biomedical ontologies already make cross-references from its concepts to concepts from other ontologies [KTM15]. For example, HPO, an ontology of human phenotypes, makes use of concepts from other ontologies to define their own concepts: *e. g.* the definition of Asymmetry of the mouth makes reference to the concepts Asymmetric from PATO and Mouth from FMA. These links, rather than the ones between equivalent concepts, are likely to be useful in my endeavour.

3.7 Recent advances

As a community, we are now empowered with tools that allow us to compare concepts and annotated entities, using methods known to work in different scenarios. Consequently, the amount of new semantic similarity measures being proposed has been decreasing. Instead, there have been different types of advances in this field of research. Recent literature tries to *i*) systematise and organise this field, so that future research can be even more powerful; and *ii*) use the already existing measures in ways that can improve their already high performance.

Harispe et al. [Har14] present a framework that tries to contribute to the understanding of semantic measures by unifying the existing measures into a single theory of semantic similarity, citing at least 21 different concept-wise measures, proposed from 1989 to 2012, and how they fit into their framework.

In one of my contributions (see Section 5.1 "Disjointness axioms in semantic similarity"), I have also proposed and validated a new measure of shared information content to use in

the chemical domain [FHC13]. Again, this is not a new measure of similarity but a *plug-in* that can be incorporated in existing measures in order to take into account different ontology constructions (in this case, disjointness axioms).

This document falls under the scope of these new research endeavours, since, as we will see in Chapter 7 "Multi-domain semantic measures", I do not propose a mutli-domain measure of similarity from scratch, but build it as a set of extensions that are based on already existing measures enabling their application on entities annotated with concepts from more than one ontology.

3.8 Summary and classification

The current measures of semantic similarity can be classified according to four axes:

Extension Measures that use only the information contained in the ontologies representing the concepts being compared are *intrinsic*, while the ones that use external resources are *extrinsic*.

Source of semantics Measures can be *edge-based* or *node-based*. They can also use other attributes of a concept, namely their labels or synonyms, which are *lexical* sources of semantics. Using information content measures to calculate the specificity of a concept is also possible for node-based measures.

Ontology multiplicity Measures can be *single-ontology*, *multiple-ontology single-domain* and *multi-domain*. Given the absence of multi-domain measures in the literature, only single-ontology and multiple-ontology single-domain measures have been described.

Aggregation technique Group-wise measures that are based on concept-wise measures must use a technique to aggregate the similarity matrix into a single value. The techniques explored in this chapter are the *average* of the matrix, the *maximum* and the *best match average* (BMA). Some group-wise measures are not based on concept-wise measures (*e. g.* sim_{UI} and sim_{GIC}).

As a summary of this whole chapter, **Table 3.1** classifies all the mentioned measures based on these four axes.

Table 3.1 – Summary of the characteristics of some semantic similarity measures. Node based measures that use the notion of information content to calculate specificity are marked with ^{IC}. Concept-wise measures do not have an aggregation technique and, as such, are marked with a dash (–) in that column. SO: single ontology; MO: Multi-ontology single-domain; BMA: best match average; GW: Group-wise measure not based on a concept-wise measure.

Publication	**The four axes of classification**			
	Extension	Source of semantics	Ontology multiplicity	Aggregation technique
[Rad89]	intrinsic	edges	SO	–
[WP94]	intrinsic	edges	SO	–
[Res95]	extrinsic	nodes^{IC}	SO	–
[JC97]	extrinsic	nodes^{IC}	SO	–
[Lor03]	extrinsic	nodes^{IC}	SO	average
[RE03]	intrinsic	nodes & lexical	MO	–
[SVH04]	intrinsic	nodes^{IC}	SO	–
[CSV05]	intrinsic	edges & nodes^{IC}	SO	adapted BMA
[LD06]	intrinsic	edges	SO	various
[Gen07]	intrinsic	nodes	SO	GW
[CSC07]	extrinsic	nodes^{IC}	SO	–
[Wan07]	intrinsic	nodes	SO	BMA
[Pes08]	extrinsic	nodes^{IC}	SO	GW
[MN09]	intrinsic	edges	MO	–
[Köh09]	extrinsic	nodes^{IC}	SO	BMA
[SLA10]	extrinsic	nodes^{IC}	SO	BMA
[FC10]	extrinsic	nodes^{IC}	SO	–
[Gre10]	extrinsic	nodes^{IC}	SO	BMA
[CS11]	intrinsic & extrinsic	nodes^{IC}	SO	–
[SBI11]	intrinsic	nodes^{IC}	SO	–
[Bat12]	extrinsic	nodes^{IC}	SO	–
[Sán12b]	intrinsic	nodes	SO	–
[Sán12a]	intrinsic	nodes	MO	–
[SB13]	intrinsic & extrinsic	nodes^{IC}	MO	–

PART II

Contributions

> Scientists have become the bearers of the torch of discovery in our quest for knowledge.
>
> — Stephen Hawking

Chapter 4

Validation strategies

One vital step in the development of semantic similarity algorithms is their validation. This step assesses the accuracy of the proposed measure with respect to a predefined goal. As such, choosing the correct validation strategy is vital to ensure the scientific soundness of the results obtained with semantic similarity: a biased or inappropriate validation strategy can erroneously certify the similarity measure and, in the worst case, lead to the validation of wrong conclusions and incorrect facts.

However, despite the importance of this step, validation of semantic similarity measures is usually carried out in an *ad-hoc* manner, with no systematization having ever been conducted around this subject. Such a systematization is important to all intervening parties (developers of semantic similarity, users of these measures, and scientific literature publishers) because:

- it provides semantic similarity developers a way to choose a validation strategy that is appropriate for their measure and its application end-goals;
- it exposes the differences and resemblances between validation strategies, thus enabling developers to choose one that is orthogonal to the ones already executed;
- it empowers users to more quickly ascertain whether the validation strategy that was used to evaluate a measure is relevant for their use cases and, by extension, whether the measure itself is appropriate for their goals; and
- it allows a standardisation of validation strategies: the existence of a controlled vocabulary that encodes the domain of validation strategies enables both developers and literature publishers to annotate their works with the classes within this hierarchy, enhancing the accuracy of metadata associated with publications.

The last item above is in accordance with the practices of semantic web (see Section 2.4 "Semantic web"), and can one day allow techniques such as semantic similarity itself to be applied on scientific literature as much as it is currently on scientific data, hence contributing to data mining and to information retrieval in general.

Realising these advantages and the lack of a proper classification of validation strategies for semantic similarity measures, I decided to contribute by assessing which strategies have been

reported in literature. During the work I carried out for this PhD, I came across a vast collection of semantic similarity measures proposed in several contexts and, as such, I am acquainted with the many strategies used to validate them. However, a scientific systematization rests on a reproducible methodology that can be carried out by anyone, irrespective of their past experience in the area. As such, I developed a method for systematically classifying validation strategies based on a literature review.

The first step was to narrow the whole semantic similarity domain. Given the popularity of GO, semantic similarity measures have been extensively proposed and studied using this ontology as a source of knowledge, and only a few works have been published that propose semantic similarity in other biomedical ontologies. As such, the systematization of validation strategies was done based on GO semantic similarity alone. This decreased the amount of literature that had to be checked but did not significantly reduce the amount of validation strategies that have been found. In this sense, the hierarchy that was created is generic on the domain of application, but contains at the moment only validation strategies found with GO-based measures. In theory, the strategies that I encountered in the literature review can be adapted and followed to validate semantic similarity in other ontologies — for example, I have previously validated semantic similarity in ChEBI by comparing it with structural similarity [FC10; FHC13]; HPO similarity has been validated by determining whether the measure can predict diseases based on phenotypes [Köh09]. Several facts contributed to my choosing GO over the other ontologies:

- It was one of the first biomedical ontologies to have been used in ontology-based semantic similarity measures [Lor03], and has since been extensively used with this purpose throughout the years (it is probably *the most* extensively used).
- It is also a formal ontology. GO is written in both OBO and OWL; therefore, it uses the first-order logic constructions that these languages provide to represent knowledge. This is in contrast with other highly used vocabularies which use instead generic and underspecified properties between concepts, such as MeSH or SNOMED-CT (see some examples in Section 2.2 "Ontologies").
- GO is in an advanced stage of development. It was the first biomedical ontology to have been developed with an objectively defined domain rather than being a general purpose vocabulary, and it is used extensively amongst the computational biology and bioinformatics communities to annotate gene products (proteins and other molecules derived from DNA that serve a function in the cell).
- Similarity measures between proteins have many different applications, including *i*) transferring knowledge between proteins [Tao07] (*e. g.* by comparing a protein with other proteins, one can predict unknown functions by hypothesising that similar proteins have similar functions); *ii*) predicting whether two proteins interact [AB04; Guo06] (either physically, by forming a complex, or in less obvious ways, like being part of the same metabolic pathway); and *iii*) automatically categorising a collection

of proteins in meaningful groups to facilitate future research in finding proteins of interest [DS05]. While protein similarity has been traditionally performed by resorting to their amino-acid sequence, with methods such as BLAST [Alt97] and the Smith-Waterman algorithm [SW81] being associated with highly relevant results in knowledge transfer [WJB04], semantic similarity has become a tool of its own, having contributed to the aforementioned tasks.

4.1 Methodology

Having selected GO as the target ontology for this classification endeavour, I developed a reproducible methodology:

1. On May 21st 2015, I conducted a search using the PubMed bibliographic database with the query `"semantic similarity" "gene ontology"`. This resulted in 121 articles being retrieved.
2. An empty set of validation strategy classes was initialised.
3. For each of the 121 articles, I read the abstract and extracted from it the validation strategies that were followed. When the abstract was insufficient to perform this step, I read instead the full text, when available.
4. Each validation strategy was classified under one of the classes in the set or, if no appropriate class existed, a new one was created.
5. Finally, the classes found in the previous step were organised in a hierarchical structure.

Step 4 is the most relevant for this task, but it is also susceptible to some subjectivity, as the classification of the measures may not always be straightforward. For example, a new strategy may be slightly different from a previously encountered one, and it is not always obvious if a new class should be created or if the two strategies should instead be classified with the same class. To minimise this subjectivity, whenever a new class is inserted in the set for this reason (*i. e.* when a more specific version of an existing strategy is found), all the strategies previously classified under that class were reassessed.

Only papers that validated semantic similarity were considered. Hence, I filtered papers that use semantic similarity as part of another system, whose main purpose is *not* the comparison of gene products, *e. g.* papers that introduce a methodology that uses semantic similarity to support protein-protein interaction queries [GVC13] or to find patterns in genome-wide associated studies [KKK13].

4.2 A hierarchy of validation strategies

The main result of this task was the hierarchy created during the literature review process. **Figure 4.1** summarises the strategies that were found by following the methodology above. The hierarchy classifies validation strategies into four main branches (represented by a grey shade in the figure), each one further divided into more specific types of validation strategies. The numbers on the right indicate the amount of papers that use a validation strategy of that type, both directly and indirectly (for example, no strategy was classified directly as "Contextual behaviour", but instead I classified papers with the leaves under that branch).

The validation strategy hierarchy produced in this task contains four branches, which can be defined as follows:

Comparison strategies The semantic similarity measure is compared to another similarity measure, which I name the *anchor*. This comparison is supported by a dataset that includes the *anchor* similarity values for pairs of proteins.

Classification strategies The semantic similarity measure is used as the basis for a classification model (through machine-learning algorithms) which is trained to predict a certain property of gene-related entities (*e. g.* proteins or pairs of proteins).

Contextual validation Semantic similarity is calculated for two kinds of protein pairs, which are hypothesised *a priori* to exhibit different similarity patterns (*e. g.* proteins that interact with one another should show higher similarity values than random protein pairs). Statistical methods are used to show that similarity in one of the groups is indeed, on average, higher than in the other group.

Theoretical validation The semantic similarity values alone are used as validation, providing a "sanity check" over the semantic similarity measure, rather than an actual validation using real-world data.

The next subsections detail the various strategies included in the hierarchy.

4.2.1 Comparison with other measures

One of the most straightforward ways to determine the performance of a semantic similarity measure is to compare it with an anchor measure to determine how well semantic similarity reflects the anchor (*e. g.* by determining the Pearson's correlation coefficient between the two measures). There are two main scenarios where it is desirable to apply this strategy:

– The anchor measure may take a long time to perform and may not be scalable for large and rapidly changing datasets. This is the case of manual similarity values assigned to pairs of proteins by experts: manual comparison is not practical for real-time systems, such as the search functionality in a protein database.

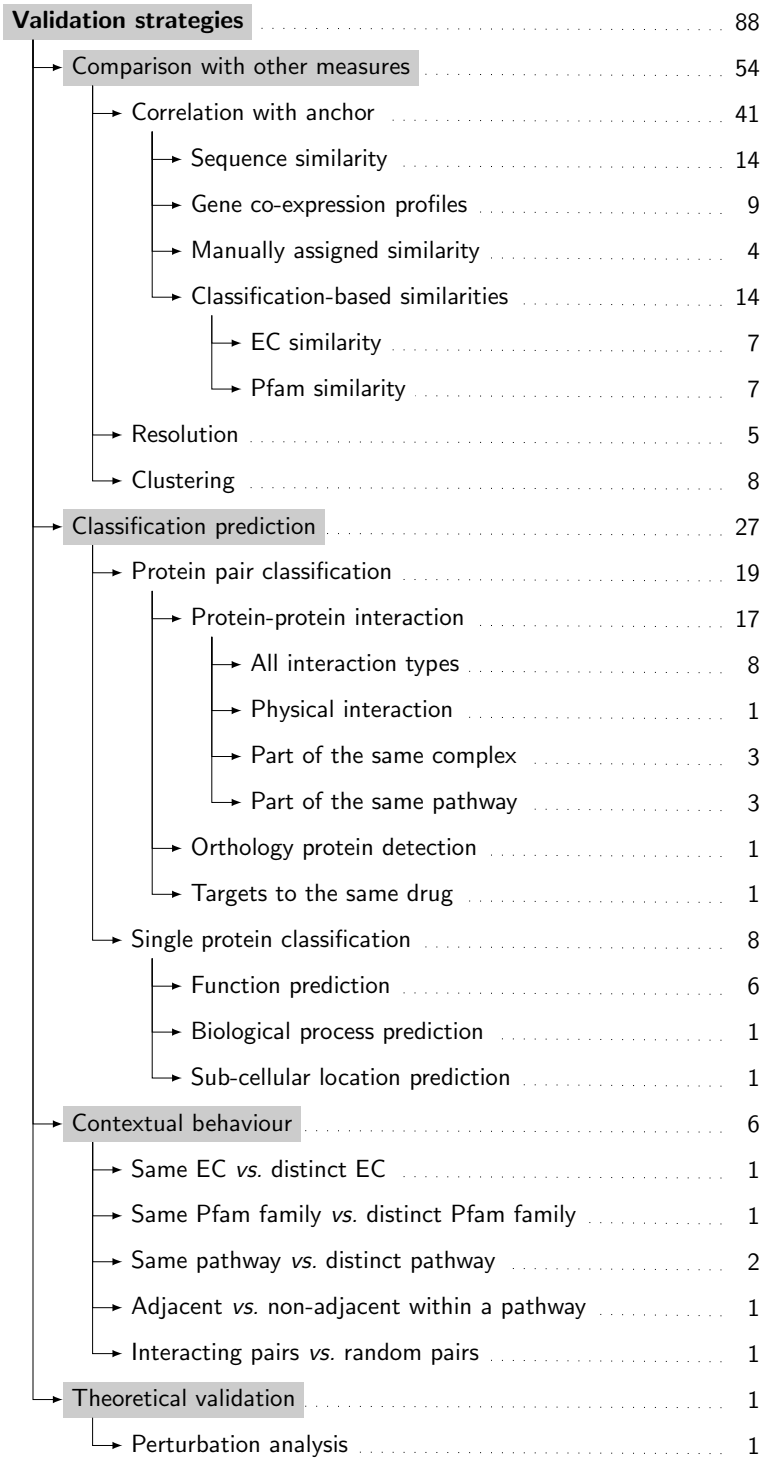


Figure 4.1 – Hierarchy of strategies employed in GO-based similarity validation. The column on the right contains the number of strategies found in literature that were classified with the corresponding class, either directly or indirectly. The only non-leaf strategy that has been used to classify a strategy is Protein-protein interaction, which correspond to works that use unspecified types of protein-protein interaction.

– The anchor measure may have already been proven successful for a certain task. In this case, deploying a new measure that highly correlates with the old one provides a good argument in favour of the suitability of the semantic measure at least in the same task. For example, sequence similarity can be used to predict protein sub-cellular localization [NR02]. Furthermore, a single similarity measure that highly correlates with several anchor measures, each developed to suit a specific task, can be regarded as a generalization of those measures.

A note of warning is needed when considering these strategies: perfect correlation is not the end goal. In fact, devising a new measure that is completely aligned with an old one can be considered a mere academic exercise, as no information can be extracted from the new measure that could not have been inferred from the anchor. The only advantage is if the new measure takes less time or less memory to compute, or if it does not depend on extra knowledge sources; usually, however, semantic similarity is slow (compared to other algorithms) and uses external information in its intermediate calculations.

Additionally, automatic anchor measures are usually known to have some shortcomings. For example, sequence similarity has been notoriously used for a few decades under the assumption that similar amino-acid sequences often correspond to similar functions [*e. g.* BK98]; but that assumption fails in some cases, such as similar sequences corresponding to disparate functions, or similar functions being performed by proteins with completely different sequences [WL03; WLT05]. As such, it is important to clearly state that, even though a high correlation with an anchor measure is an argument for the suitability of the new measure, care must be taken when interpreting actual correlation coefficients between semantic and other non-manual similarity measures.

Correlation strategies differ essentially on the anchor measure they use:

Sequence similarity measures These assign a numeric value to a pair of proteins based on their amino-acid sequences. Sequence similarity measure used in these validation strategies are based on Smith-Waterman [SW81] and BLAST [Alt97].

Gene co-expression profiles The expression levels of two genes are compared in several different situations, and the absolute value of Pearson's correlation coefficient between the expression values throughout these situations is measured. Based on the assumption that similar genes exhibit similar expression levels in the same situation (*i. e.* they present a high overlap in their expression profiles), a high correlation between their expression profiles and the semantic similarity between the pair is used to validate the measure [*e. g.* Wan04; JB10; YNP12].

Manual similarity Sometimes, it is possible to have an expert go through a series of pairs of protein or pairs of $\mathbb{GO}$ concepts and assign each one a similarity value, which the semantic similarity measure must reflect [*e. g.* Xu13].

Classification-based similarity Automatic similarity derived from the manual classification of the proteins has also been used. One example is the use of the Enzyme

Commission (EC) classification [Mos15] to compare two enzymes (the similarity is the number of levels in the two EC numbers that match); this strategy was introduced by Pesquita et al. [Pes09a]. Another example is the Pfam classification [Bat02] (similarity is the number of shared families between the two proteins); this validation strategy was introduced by Couto et al. [CSC07].

Another way to determine the performance of a semantic similarity measure is to calculate its *resolution* with respect to the anchor measure, a numeric value that reflects the overall behaviour of the measure. This evaluation was first introduced by Pesquita et al. [Pes08], and is defined as "the relative intensity with which (on average) variations in the sequence similarity scale are translated into the semantic similarity scale". The assumption is that the higher the resolution, the more accurate is the semantic similarity measure, as it can reflect small differences in two proteins that a measure with less resolution cannot.

The final strategy in this branch consists in using the semantic similarity measure to cluster proteins and then compare the resulting cluster with a reference, which may be manually assembled or can itself be based on other resources. Wang et al. [Wan07] validate semantic similarity by manually assessing whether the results of hierarchical clustering reflect an expert notion of clustering. Automatic alternatives to this include the Davies-Bouldin Index [DB79] or the Fowlkes-Mallows Index [FM83], which measure the degree to which two clustering results overlap.

4.2.2 Classification strategies

Semantic similarity can also be validated by assessing whether it can predict properties of proteins or protein pairs. For these strategies, a dataset of known property values must be given in advance (the "gold-standard"), and the validation strategy usually consists in determining some kind of accuracy of the semantic similarity measure in predicting these properties.

4.2.2.1 Protein-protein interactions

The most straightforward way of basing classification problems on semantic similarity values is to use the similarity between two proteins to predict whether they interact. Interaction, in this context, can be interpreted as:

- the actual physical, momentary interaction between the proteins, such as when one of the proteins modifies the other protein (*e. g.* through phosphorylation);
- the long-lived interaction between proteins that are part of the same multi-protein cluster (*e. g.* ribosomes consist of a series of proteins and, thus, form a multi-protein complex); and
- a more abstract notion of interaction that occurs when the two proteins are part of the same metabolic pathway (*e. g.* they both regulate the same process).

In these classification problems, a dataset of positive pairs is provided containing pairs

of proteins that are known to interact. Negative pairs can be gathered from the literature (*e. g.* from journals such as the Journal of Negative Results in Biomedicine) or randomly generated. Furthermore, random generation can be *i*) blind, *i. e.* any pair is accepted in the set; or *ii*) generated in such a way that it is known to contain few positive pairs. This last method can drastically reduce the chances that a positive pair ends up in the negative dataset. For example, since proteins that are part of the same cluster must necessarily coexist in the same cellular location, the generation of negative pairs may exclusively generate pairs of proteins that are known to be located in separate cell compartments. However, Ben-Hur and Noble [BN06] argue that building the negative set in this way can lead to unreliable performance indicators, because there are proteins in the same cellular compartment that are not, in fact, part of the same complex. Thus, this selection method introduces a bias.

Table 4.1 describes some of the validation strategies that were found in the literature review and that were classified under "Protein-protein interaction". All the strategies use some sort of online biomedical database to create the positive dataset, while generally the negative dataset is randomly constructed. Several types of interaction can be used, even within the same strategy, and the performance is usually reported *i*) as some statistical test (for example, the p-value associated with the capacity of the measure to predict the correct interactions), *ii*) as the value of precision, or *iii*) as the value of the Area Under the Curve (AUC) of a Receiver Operating Characteristic curve [Faw04; Faw06].

Several datasets typically used by these strategies include:

- KEGG, the Kyoto Encyclopedia of Genes and Genomes [Oga99], can be used to find proteins that participate in the same pathway or that are part of the same protein complex;
- CORUM [Rue08] is a dataset of mammal protein complexes; and
- DIP [Sal04] is an all-purpose interaction database, containing at least 28 different protein-protein interaction types.

4.2.2.2 Orthology detection

Another property of protein pairs that can be predicted by using semantic similarity is the implicit property that exists between orthologous proteins. The main assumption of this strategy is that orthologs (*i. e.* proteins whose genes are, in evolutionary terms, descendent of the same ancestral DNA sequence) should exhibit a higher similarity than other pairs of proteins. Wu et al. [Wu13] introduced this idea and used a statistical test to validate semantic similarity by noticing that the similarity values between orthologous proteins is higher than between random protein pairs.

Table 4.1 – Protein-protein interaction validation strategies. This table shows the details of representative instances of validation strategies based on protein-protein interaction prediction. Each strategy can have more than one data source to construct the positive and negative pairs of the dataset, as well as using multiple interaction types and performance measures.

Paper	Dataset		Interaction type	Performance
	positive pairs	*negative pairs*		
[AB04]	custom dataset	custom dataset	same complex	statistical test
[Guo06]	KEGG PATHWAY, KEGG MODULE, BIND	random	same pathway, same complex	AUC, statistical test, precision
[JB10]	DIP	random	physical, same pathway	AUC
[MD12]	KEGG PATHWAY	KEGG PATHWAY	same pathway	statistical test
[YNP12]	CORUM	random	same complex	AUC
[Vaf13]	I2D, Reactome KEGG, NetPath, NCI-PID, CORUM	random	physical, same pathway, same complex	AUC

4.2.2.3 Single protein property prediction

While classification problems involving protein pairs are the most common, there are validation strategies directed at the properties of individual proteins as well. Three such examples have been found in the literature: prediction of protein function, prediction of the biological processes in which the protein participates, and prediction of sub-cellular localization. These three prediction problems map directly to the three GO branches: molecular function, biological process and cellular component. In fact, these strategies can be rephrased as follows:

> “Given a set of GO annotations, predict new annotations to go along with them.”

This is a way of *enriching* an annotation set with more concepts.

In these strategies, semantic similarity between the protein whose property is being predicted and the proteins whose property value is already known is used as part of a machine-learning strategy. As such, the training dataset must already contain the known property values: *e. g.* the training dataset for the function prediction problem must contain a set of proteins and their actual function(s). Performance is reported using measures frequently used in machine-learning, such as the accuracy of the machine-learning algorithms, or the AUC of the corresponding ROC curve.

4.2.3 Contextual behaviour

Like the strategies of the previous branch, "contextual behaviour" strategies are based on the assumption that proteins that are in some way related should exhibit a higher similarity (in general) than the ones where that relation does not hold. However, instead of predicting properties of proteins, these methods consist in only observing whether that assumption holds. For example, proteins that are adjacent in a certain metabolic pathway should exhibit higher average semantic similarity than proteins of that pathway that are not adjacent.

In order to prove that this behaviour holds, statistical methods are frequently used to show that average semantic similarity in one of the groups is statistically higher in the other group, which can be achieved using, *e. g.* Z-test or Student's t-test [Ros10].

Strategies found in the literature search include dividing protein pairs depending on whether the two proteins:

- have the same EC classification;
- have the same Pfam family;
- participate in the same metabolic pathway;
- are adjacent or non-adjacent within a metabolic pathway; or
- form a known protein-protein interaction.

It must be noted that these strategies are not technically much different from the use of an anchor measure. For example, we could create an anchor measure that assigns 1 to protein pairs where the two proteins have the same EC classification and 0 to other protein pairs, and then correlate this measure to semantic similarity. However, anchor measures tend to be continuous rather than categorical (they return a real number in a range, usually between 0 and 1). Furthermore, in contextual behaviour strategies, the goal is to determine whether we can observe a statistically significant difference between the semantic similarity in one group with respect to another group, rather than to calculate a correlation coefficient.

4.2.4 Theoretical validation

Theoretical validation strategies depend only on the actual semantic similarity values between pairs of proteins and not on any other information about those pairs.

The only validation instance I encountered in this branch was the calculation of the "resistance to ontology perturbation" [MD12]. This strategy measures how much the semantic similarity values change when the ontology underlying the measure is changed. Being robust to perturbation is regarded as a necessary condition for a useful similarity measure, as ontologies change over time and the (usually) small variations from one version to the next should not have a significant impact on the similarity values calculated between proteins. For a robust measure, increasing perturbation rates cause increasing deviation. In a non-robust measure, deviation does not correlate with perturbation rates.

4.3 Results

While the previous section contains a detailed description of the proposed hierarchy, this section describes the general results obtained from the review process.

Of the 121 papers retrieved from PubMed, 45 provide one or more validation strategies for semantic similarity measures, for a total of 88 distinct strategies. The most frequently used strategies are "Comparisons with other measures" and "Classification predictions", which together amount to more than 90% of the strategies found.

Another result obtained with this literature review (which is absent from the hierarchy) is that gene products other than proteins are never explicitly mentioned in any of the strategies and in fact, they are not addressed in most of the reviewed papers. This seems to suggest that most semantic similarity measures in $\mathbb{GO}$ are developed and applied to proteins only.

Another result not represented in the hierarchy is related to the papers that did not contain semantic similarity validation strategies. I found 44 papers that use semantic similarity as part of another system, 8 that use semantic similarity to validate other techniques, and 6 that use semantic similarity to find new knowledge, such as the identification of transcription factors involved in some cellular response [Sek15]. These papers assume that the semantic similarity measures they use are valid for their purpose.

The rest of the papers (no validation strategy and no assumption on the validity either) are distributed as follows:

- 5 papers present and provide software to compute semantic similarity (2 web-based tools, 2 R packages and 1 desktop application);
- 4 papers are theoretically oriented (they present mathematical or statistical frameworks on top of the existing semantic similarity measures);
- 3 papers mention semantic similarity but do not propose new measures nor do they validate existing ones;
- 2 papers provide a database of pre-computed semantic similarity values;
- 1 is a review of semantic similarity;
- 1 uses semantic similarity outside of $\mathbb{GO}$;
- 1 has been retracted; and finally
- 1 does not provide enough information in the full text to classify its validation strategies.

4.4 Discussion

This hierarchy is meant to be used *i*) by semantic similarity developers when assessing the validity of their measures; *ii*) by general researchers, as it facilitates the process of selecting a semantic similarity measure based on whether it has been validated with a strategy that

overlaps their needs; and *iii*) by literature authors, since it allows them to contextualise their work under a controlled vocabulary of validation strategies, thus enabling its easy replication and the integration of their results in other research.

It is of practical relevance, therefore, to expose some of the advantages and disadvantages of the validation strategies, at least at the high level of the four branches of the hierarchy (these features are summarised in **Table 4.2**):

Comparison strategies These methods are often easy to implement, and provide a general idea of the behaviour of the measure in the full spectrum of similarity, since they can be readily applied to any pair of proteins as long as the anchor measure can be calculated or is known. As discussed previously, a perfect correlation means that the measure is exactly equivalent to the anchor measure, and thus cannot provide any new information that the anchor measure does not already provide. As such, for high values of correlation, a higher correlation does not necessarily correspond to a more useful measure.

Classification strategies The main advantage of these strategies is that they provide a practical, real-world-based evaluation, since they actually answer a relevant question: "Can my measure be used to predict X?". However, at least three disadvantages exist. First, they require a large dataset (the gold-standard), which is not always available. Second, choosing the appropriate machine-learning algorithm is hard and strictly depends on the data. For example, while most works classify a pair of proteins as positive if their semantic similarity is above a threshold, single-protein classification cannot directly employ this idea. Finally, there is a bias associated with the choice of training dataset: while the semantic similarity measure being validated may be able to properly classify the instances in the gold-standard, it may not perform so well in other data.

Contextual behaviour strategies Like classification strategies, these strategies require a dataset that contains protein pairs along with some annotation (*e. g.* they are part of the same pathway, or physically interact); unlike those strategies, however, comparing the average semantic similarity values in the two groups is simple and usually resorts to sound statistical methods.

Theoretical strategies These strategies can be used to check properties of the proposed measure (*e. g.* mathematical, statistical or behavioural properties, such as the triangle inequality) but may otherwise have no external significance.

Given these features, I developed a pipeline to help semantic similarity developers choose the most appropriate validation strategies for their measure (see **Figure 4.2** on page 59). Since classification strategies are the ones with more practical applications, these types of validation strategies should be selected whenever possible, followed by contextual behaviour strategies, then comparison strategies and finally theoretical strategies. A stronger validation assessment, however,

Table 4.2 – Features of the several types of validation strategies. Each row contains an advantage and each column represents one of the four branches of the hierarchy. The × sign marks the presence of the advantage for the strategy type and the ? represents absence of enough information (small number of examples found in the literature) to enable generalisation of the feature.

	Comparison	**Classification**	**Behaviour**	**Theoretical**
Real-world application		×		
Independent of external data	×			×
External significance	×	×	×	
Easy to implement	×		×	?

makes use of more than one strategy type, and as such the diagram does not terminate when a strategy type has been selected but continues down the order specified above (*cf.* the dashed lines in the image). For example, a developer that can perform a classification strategy should, nevertheless, if possible, try to correlate their measure with anchor measures as well. Additionally, whenever it is important that the measure satisfies mathematical and/or statistical properties, theoretical validation strategies should be followed. For example, Chow and Rodgers [CR05] describe a method to draw Venn diagrams where the areas of the intersections are proportional to the amount of overlap between the groups and which requires the triangle inequality to hold.

Finally, a concluding remark on the hierarchy itself is that it is not comprehensive in at least two senses:

- More specific validation strategies than the ones included in this review can be inserted into the hierarchy (either in one of the already existing branches or directly below the root of the hierarchy). Indeed, future research may require that the hierarchy be updated. For example, I decided not to subdivide the strategy "Correlation with gene co-expression profiles", since the 9 instances found are all essentially equivalent. In the future, however, if a validation strategy uses a more specific version of this methodology (*e. g.* by restricting the situations used to compute the co-expression profile), new classes should be added. Additionally, I tried to make the classes in the hierarchy as distinct as possible, but do not guarantee actual disjointness between them.
- Although the literature search is representative of the space of validation strategies followed in $\mathbb{GO}$-based semantic similarity measures, the search query does not exhaustively find all relevant documents. For example, some papers use the expression "functional similarity" instead of "semantic similarity" and thus were not found.

4.5 Conclusions

The task presented here consisted in a systematic review of the strategies used to validate GO-based semantic similarity measures. My review resulted in the development of a hierarchy of validation strategies, which encompasses, to the best of my knowledge, most of the strategies applied so far in this domain. The most frequently used strategies are the comparison with other similarity measures and the use of semantic similarity for predicting protein-protein interactions.

In the future, I intend to work on a tool that assists interested semantic similarity developers in setting up a validation step, akin to an already existing system developed for that effect (the Collaborative Evaluation of GO-based Semantic Similarity Measures [Pes09a]), which will *i*) provide automatic ways to download datasets and GO annotations, *ii*) ask the user to supply the similarity values for the necessary protein pairs, and *iii*) perform the computations necessary to validate the measure, according to user-selected strategies.

Additionally, as part of the efforts in biomedical research, I foresee the possibility to encode this hierarchy into an actual ontology, which other users can reference and use to annotate their papers. For example, this hierarchy can be included under the concept Validation from the Ontology for Biomedical Investigations (OBI), an ontology frequently used by the biomedical informatics community to annotate experimental protocols. Since other domains of research that make use of semantic similarity also require the similarity methods to be validated, I anticipate that this hierarchy will be useful to these domains.

Even though many validation strategies followed outside the scope of GO already fit into the hierarchy not all of them do. For example, protein-protein interaction is a methodology specific to proteins and, consequently, does not map to the other domains. As such, extension of the hierarchy to accommodate other domains is also part of my future plans.

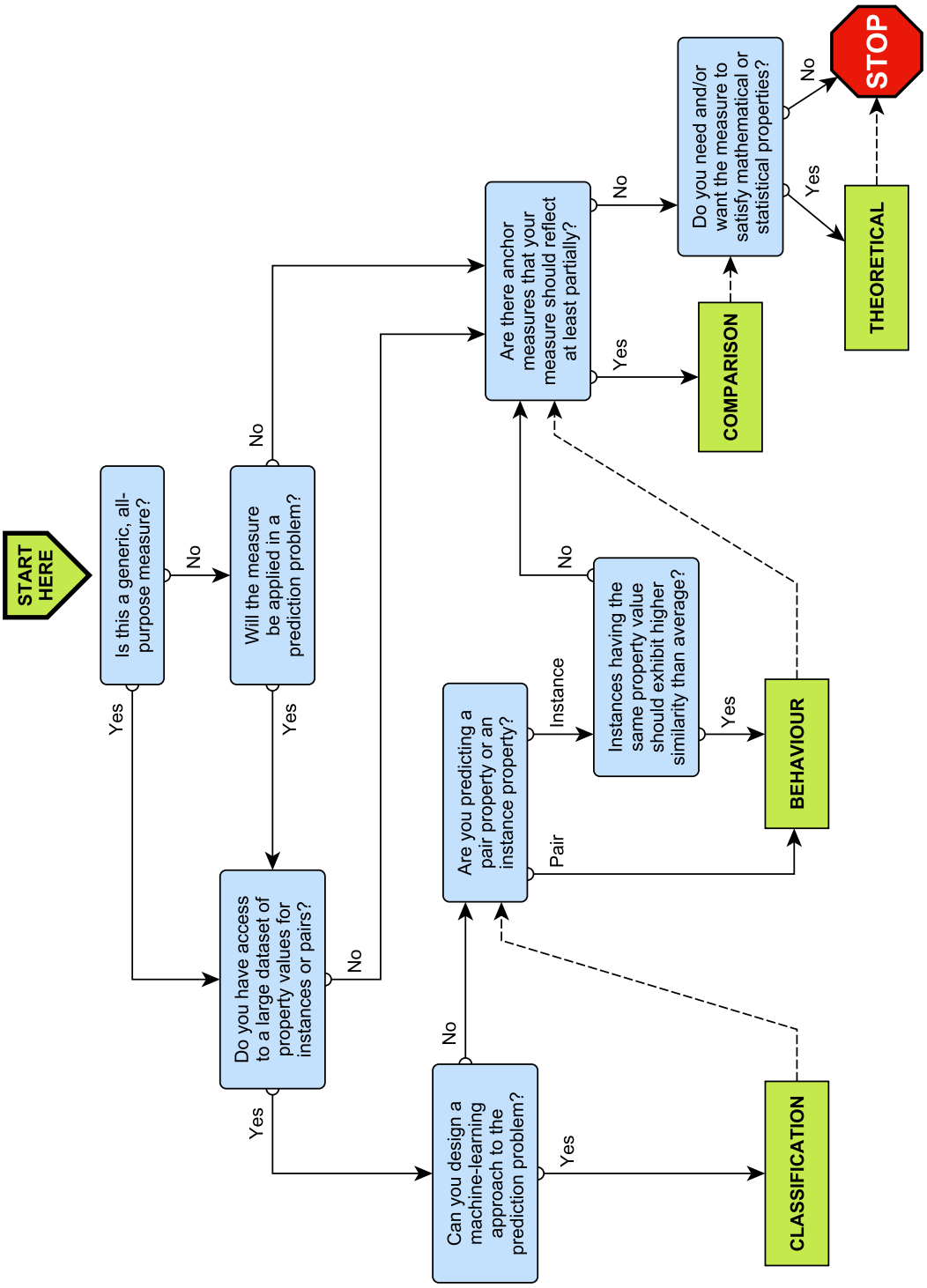


Figure 4.2 – Pipeline to assist semantic similarity developers in the validation step. Specific classes should be selected according to the data that the developers have access to, and to the overall goal of the measure. Questions the developers must answer are in blue rounded rectangles; strategy types are in straight green rectangles. The diagram anticipates the possibility of simultaneous validation strategies by drawing arrows from the resulting strategy types to additional questions (dashed arrows).

Chapter 5

Towards OWL-aware similarity

The biomedical informatics community is actively committed to the adoption of formal-logic knowledge representation languages, such as the Web Ontology Language (OWL), and the use of reference ontologies to annotate biomedical resources (*cf.* Appendix A "List of ontologies"). This adoption has resulted in the increase of *i*) the amount of knowledge represented in ontologies, and *ii*) the quality of these representations in respect to the reality.

As has been argued by Couto and Pinto [CP13], there are some benefits to considering exploiting formal axioms in the calculation of semantic similarity. In this chapter, I report the enhancements that I achieved in the pursuit of semantic similarity measures that can maximise the use of these axioms. I first report on a measure that can use disjointness axioms and then on another that can use existential quantifications.

5.1 Disjointness axioms in semantic similarity

5.1.1 The idea

Disjointness axioms are one example of formal constructions in OWL ontologies that are gaining momentum in the biomedical informatics community. For example, the Chemical Entities of Biological Interest (ℂ𝕙𝔼𝔹𝕀) ontology now includes this type of axioms [Has13]. To explore this new type of information, I devised an algorithm that can be *plugged* into some semantic similarity measures to take into account disjointness information.

A disjointness axiom declared for a pair of concepts express the constraint that an instance of one of them cannot also be an instance of the other. This constraint logically implies that the two concepts cannot have common subclasses. Should such shared subclasses be detected by a reasoner, the reasoner will flag the ontology as *inconsistent* [SSL13], which can be used by ontology developers to prevent errors in ontology development. For example, in ℂ𝕙𝔼𝔹𝕀, this technique has been used to detect that a specific ion (a type of charged molecule) was misclassified as a group (a strict part of a molecule) [Has12].

Figure 5.1 illustrates this situation: this ontology snippet asserts that no instance of Rectangle can simultaneously be an instance of Trapezoid. However, given the open-world

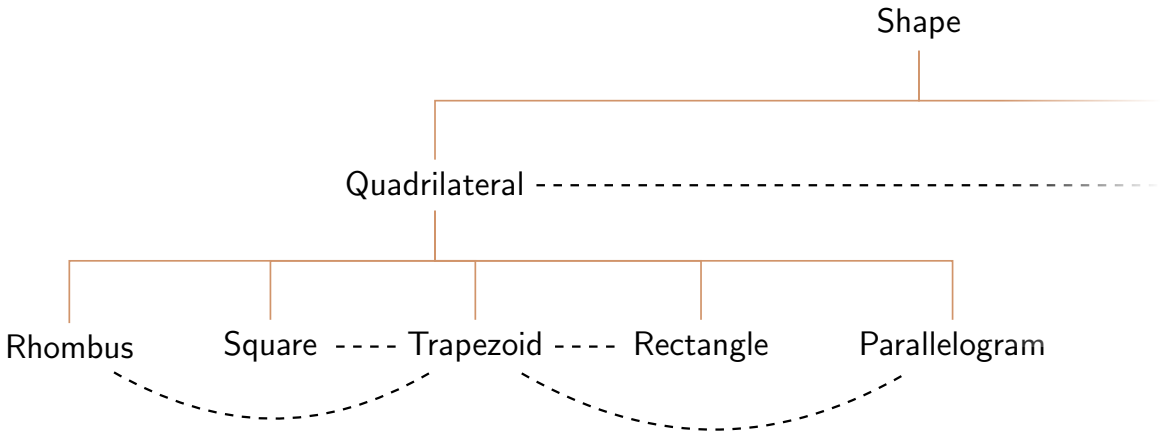


Figure 5.1 – A snippet of a hypothetical Shape Ontology. In this snippet, I use the term Parallelogram as "a quadrilateral with two pairs of parallel sides" and Trapezoid to mean "a quadrilateral with two parallel sides and two obtuse angles". Solid lines represent class-subclass relationships, dashed lines represent disjointness axioms. Note that a proper shape ontology would classify Square as a subclass of Rectangle, Rhombus and Parallelogram. For the sake of the argument being exposed, however, assume that such information is yet unknown by the ontology creators.

assumption that underlies OWL ontologies (see Section 2.2 "Ontologies"), there can be instances of Rectangle that are also instances of Parallelogram (in fact, it is a consequence of the relevant geometric definitions that squares are both rectangles and parallelograms). For this reason, the similarity between Rectangle and Parallelogram should intuitively be higher than the similarity between Rectangle and Trapezoid. Using σ to represent the function that returns the similarity between two concepts, this hypothesis can be mathematically stated with eq. 5.1:

$$\sigma(\mathsf{Rectangle}, \mathsf{Parallelogram}) > \sigma(\mathsf{Rectangle}, \mathsf{Trapezoid}). \tag{5.1}$$

5.1.2 The proposed measure

As has been discussed in Section 3.3 "Node-based approaches", there has been an effort to design measures that compute the shared information content between two concepts: while shared information content between concepts x and y has been assumed to be well estimated by the maximal information content of the concepts that subsume both x and y, Couto and Silva [CS11] suggest DiShIn, a shared information content measure that builds upon existing measures (such as the ones in eqs. 3.4 and 3.7) and which behaves as a *plug-in* to such measures. In its particular case, DiShIn explores *multiple parentage* in order to ensure that all shared information across multiple ancestors is taken into account.

Likewise, instead of developing a new semantic similarity measure to deal with disjointness axioms, I proposed a *plug-in* to be used on top of existing measures of shared information content. My *plug-in* refines the estimation of shared information between two concepts by incorporating

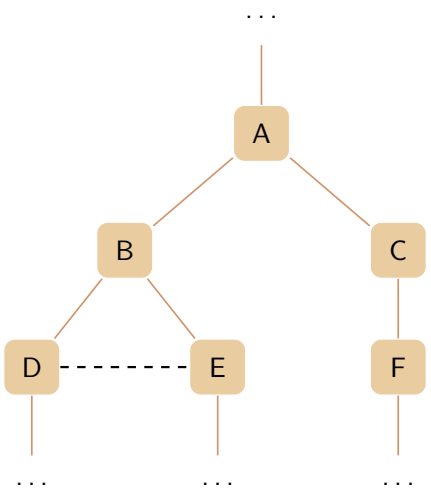


Figure 5.2 – Example ontology with disjointness axioms. This illustrates a toy ontology that shows some concepts in a class-subclass hierarchy (represented with solid lines) as well as a disjointness axiom asserted between two concepts (dashed lines).

the disjointness axioms asserted in the ontology. In here, I denote the disjointness-aware measure of shared information content between concepts x and y as $\mathrm{IC}_\mathrm{s}^\mathrm{disj}(x, y)$, which is calculated based on an existing measure of shared information content, $\mathrm{IC}_\mathrm{s}(x, y)$.

Given the example presented in **Figure 5.1** and the inequality of eq. 5.1, it would be desirable for the measure of shared information content to decrease for concepts that are known to be disjoint, formalising the intuition that disjoint concepts are less similar because they cannot share subclasses. Furthermore, to respect the open-world assumption, the measure should stay unchanged when two concepts are *not known* to be disjoint.

With these constraints in mind, I proposed the following measure of shared information content:

$$\mathrm{IC}_\mathrm{s}^\mathrm{disj}(x, y) = \mathrm{IC}_\mathrm{s}(x, y) - k(x, y) \tag{5.2}$$

where $\mathrm{IC}_\mathrm{s}(x, y)$ is any measure of shared information content between x and y, $k(x, y) > 0$ if x and y are disjoint and $k(x, y) = 0$ otherwise.

Two points were crucial in the development of this measure. First, note that, as is, this equation presents a *discontinuity*. In the hypothetical ontology of **Figure 5.2**, this measure implies

$$\mathrm{IC}_\mathrm{s}^\mathrm{disj}(\mathsf{D}, \mathsf{E}) < \mathrm{IC}(\mathsf{B}), \tag{5.3}$$

which, depending on the value $k(\mathsf{D}, \mathsf{E})$, could lead to

$$\mathrm{IC}_\mathrm{s}^\mathrm{disj}(\mathsf{D}, \mathsf{E}) < \mathrm{IC}(\mathsf{A}) = \mathrm{IC}_\mathrm{s}^\mathrm{disj}(\mathsf{D}, \mathsf{F}). \tag{5.4}$$

However, this should not be possible, since D and E share more information than D and F. Therefore, k must be bounded according to the IC of the most informative ancestor of the

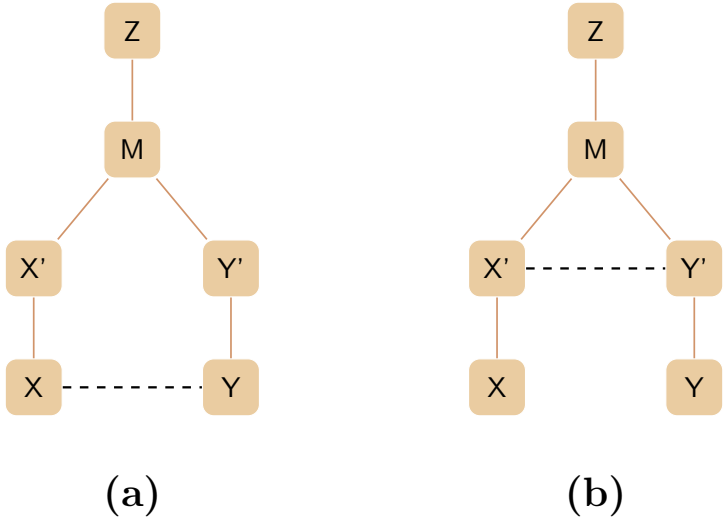


Figure 5.3 – The potential for implicit common superclasses between two concepts. In both cases, MICA(X, Y) = M, and the most informative ancestor of M is Z. The difference is in the location of the disjointness axiom. In situation A, there is a higher likelihood of implicit common ancestors (ICA) between X and Y, because the axiom of disjointness is further down from their common ancestry.

MICA, which, in this case, results in

$$k(\mathsf{D},\mathsf{E}) \leq \mathrm{IC}(\mathsf{B}) - \mathrm{IC}(\mathsf{A}). \tag{5.5}$$

The second major decision in the development of this measure is related to an operational notion. In fact, I have not yet defined how to compute $k(x, y)$. I propose an algorithm based on the notion of potential for *implicit common superclasses* (ICS), which measures the likelihood that two concepts share non-asserted superclasses. Consider the ontology snippets in **Figure 5.3**. In situation A, given the open-world assumption, there is a small chance that Y turns out to be a subclass of X′, while in situation B that cannot happen, since Y is inferred to be disjoint with X′ (the disjointness axiom is higher up in the hierarchy). This suggests that there is a higher potential for ICS between the concepts X and Y in situation A.

I model the *unlikelihood* of ICS as $f(x, y)$, a function that returns higher values for situations with lower potential for ICS:

$$f(x,y) = \max\left(\left\{\frac{1}{p(a,b)} \,\middle|\, a \in \mathrm{A}(x) \wedge b \in \mathrm{A}(y) \wedge J(a,b)\right\} \cup \{0\}\right) \tag{5.6}$$

where $\mathrm{A}(c)$ is the set of superclasses of c (including c), $J(a, b)$ is true when a and b are disjoint (either by assertion or inference) and false otherwise, and $p(a, b)$ is the edge length of the shortest path from a to b, taking into account only class-subclass relationships, not the disjointness arcs (only the solid edges in the figures, not the dashed ones).

Using the example ontologies in **Figure 5.3**, we can illustrate this definition by calculating $f(\mathsf{X},\mathsf{Y})$. In A, $J(a, b)$ is true only for $(a, b) = (\mathsf{X},\mathsf{Y})$; the shortest path from X to Y is

$\mathsf{X} \to \mathsf{X}' \to \mathsf{M} \to \mathsf{Y}' \to \mathsf{Y}$, which has length 4. Therefore,

$$f(\mathsf{X}, \mathsf{Y}) = \max\left\{\frac{1}{4}, 0\right\} = \frac{1}{4}. \tag{5.7}$$

In B, $J(a, b)$ is true for $(a, b) \in \{(\mathsf{X}, \mathsf{Y}), (\mathsf{X}, \mathsf{Y}'), (\mathsf{X}', \mathsf{Y}), (\mathsf{X}', \mathsf{Y}')\}$. These correspond to paths of length 4, 3, 3 and 2, respectively, leading to

$$f(\mathsf{X}, \mathsf{Y}) = \max\left\{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 0\right\} = \frac{1}{2}. \tag{5.8}$$

Additionally, in situation A, $f(\mathsf{X}', \mathsf{Y}') = 0$, since the two concepts are not disjoint and as such $J(a, b)$ is always false.

The general procedure to calculate $\mathrm{IC}_\mathrm{s}^\mathrm{disj}(x, y)$ is, therefore:

1. Determine $M = \mathrm{MICA}(x, y)$
2. Determine $Z = \arg\max_c\{\mathrm{IC}(c) \,|\, c \in \mathrm{A}(M)\}$, *i. e.* the most informative ancestor of M;
3. Calculate $f(x, y)$, as described in eq. 5.6;
4. Calculate $k(x, y) = f(x, y) \cdot (\mathrm{IC}(M) - \mathrm{IC}(Z))$;
5. Calculate $\mathrm{IC}_\mathrm{s}^\mathrm{disj}(x, y) = \mathrm{IC}_\mathrm{s}(x, y) - k(x, y)$.

With this procedure, the new shared information content is estimated as a weighted average between $\mathrm{IC}(M)$ and $\mathrm{IC}(Z)$, where a low potential for ICS leads to a shared information content close to $\mathrm{IC}(Z)$ and a higher potential for ICS leads to a shared information content close to $\mathrm{IC}(M)$. This means that the shared information content decreases by a larger amount when there is a smaller potential for implicit common superclasses. Note that if the two concepts are not disjoint, $k(x, y) = 0$ and $\mathrm{IC}_\mathrm{s}^\mathrm{disj} = \mathrm{IC}_\mathrm{s}$, which satisfies the open-world assumption mentioned previously.

5.1.3 Validation

According to the hierarchy presented in Chapter 4 "Validation strategies", I classify the validation approach followed in this work as a "Comparison with an anchor measure". For this purpose, I calculated structural similarity between ℂℍ𝔼𝔹𝕀 concepts (details can be found on the paper), and semantic similarity using the algorithm described above, and measured the Pearson's correlation coefficient between the two measures.

The proposed measure was validated in three steps, by measuring the increase in coefficient *i*) in the presence *vs.* absence of disjointness axioms, *ii*) with increasing fractions of the total number of disjointness axioms, and *iii*) in several random datasets.

For a fully detailed discussion of these three steps, I refer the reader to the published paper. Here I summarise the main results.

1. Increase in correlation coefficient I applied the new measure of shared information content to a subset of ℂℍ𝔼𝔹𝕀, including some disjointness axioms [Has12; Has13]. Since it consists

of a *plug-in* and relies on a previously defined measure of shared information content, I used, in this assessment, the classical notion of shared information content proposed by Resnik [Res95]:

$$\mathrm{IC_s}(x, y) = \mathrm{IC}(\mathrm{MICA}(x, y)) \tag{5.9}$$

where information content was calculated using eq. 3.3.

Given the sparsity of disjointness axioms relative to the size of ℂℍ𝔼𝔹𝕀, random pairs of concepts would rarely touch any disjointness information and it would be difficult to detect the difference resulting from these axioms. As such, the concepts in the dataset were chosen so that a significant part of them had disjointness information. Additionally, given the need to calculate structural similarity, the concepts in the dataset were also selected in order to ensure that structural similarity was possible. While these two constraints slightly bias the generated dataset, I believe that the bias did not result in a dataset too different from reality: in fact, although the amount of disjointness axioms already added to ℂℍ𝔼𝔹𝕀 is not high, true disjointness exists for most pairs of concepts.

Structural similarity was calculated using PubChem's fingerprint method [Bol08]. Structural similarity between ℂℍ𝔼𝔹𝕀 concepts that do not represent actual molecules but rather chemical groups (*e. g.* there is not a single chemical structure for Hexose but rather a collection of them) was achieved by comparing the collection of structures.

For the dataset created above, I compared all compounds with all the other compounds using three measures: structural similarity, classical $\mathrm{IC_s}$ and $\mathrm{IC_s^{disj}}$. I used Wolfe's t-Test [Wol76; Ros10] to determine the statistical significance of the increase in the correlation coefficient between the pair $(\text{structural}, \mathrm{IC_s})$ and the pair $(\text{structural}, \mathrm{IC_s^{disj}})$.

The Pearson's correlation coefficient between the structural measure and $\mathrm{IC_s}$ is 0.69883, and after taking the disjointness axioms into account, the correlation between structural similarity and $\mathrm{IC_s^{disj}}$ becomes 0.71571. This represents an increase of 0.01688. Despite the small absolute increase, this value is statistically significant, with a p-value of 4.5×10^{-8}. The small increase of the correlation can be attributed to at least three factors:

- As the annotation of disjointness is still incomplete in ℂℍ𝔼𝔹𝕀, we have access to only a small subset of all the *real* disjointness axioms that can be expressed in ℂℍ𝔼𝔹𝕀, which means that the shared information content changes only for a fraction of all the concept pairs (39% of the pairs in the dataset). I expect that, as the number of disjointness axioms added to ℂℍ𝔼𝔹𝕀 increases, both this fraction and the difference between correlation coefficients will increase.
- While highly correlated, structural similarity and semantic similarity measures are inherently different, and as such there is a maximum bound on the actual correlation that can be expected between the two (*cf.* Section 4.2.1 "Comparison with other measures").
- Disjointness is only one of the logical axiom types that are used to express concept

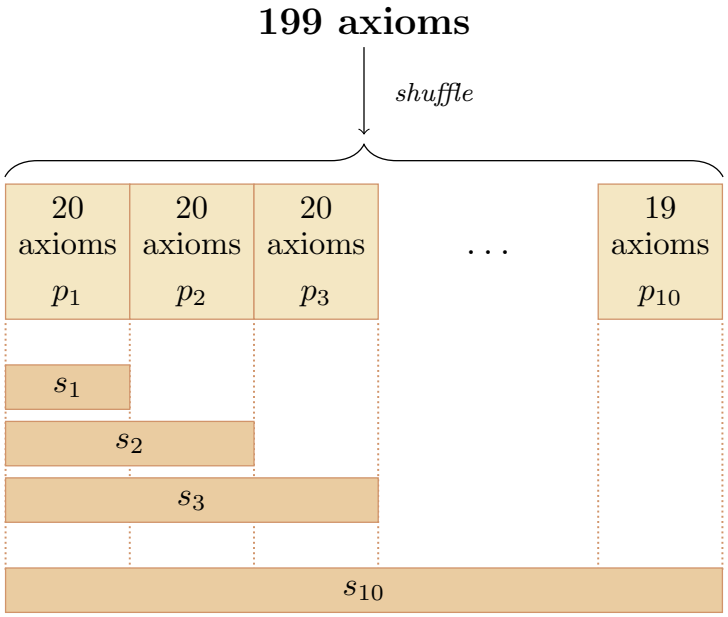


Figure 5.4 – Partitioning the set of disjointness axioms. The process used to assess the effect of the number of axioms on the correlation coefficient between structural and semantic similarity. The axioms are randomly partitioned into clusters p_1 to p_{10}. Consecutively, each of these clusters is joined with the previous ones to create the sets $s_i = \bigcup_{j=1}^{i} p_j$, which are then used to compute the increase in correlation coefficient.

definitions in an OWL ontology. In fact, ℂHEBI contains a number of other properties that are also used to capture the meaning of its concepts, *e. g.* the property *has-tautomer*, which connects together closely structurally related chemicals, and *has-role*, which connects a chemical concept to its biological activity.

2. Effect of the number of axioms The second assessment step measured the effect of the number of disjointness axioms on the correlation coefficient. I partitioned the 199 disjoitness axioms into 10 sets: the first contained 20 random axioms, the second contained these plus another 20 random axioms, *etc.*, with the final one containing the 199 axioms (see **Figure 5.4**). For each set, I calculated $\mathrm{IC}_{\mathrm{s}}^{\mathrm{disj}}$ on the previous dataset and plotted a graph showing the increase in correlation *vs.* the number of axioms. To remove any bias that resulted from the random method used to partition the axioms, I repeated this process 20 times.

This validation step has the objective of simulating the development of ℂHEBI ontology with respect to the number of disjointness axioms. For each of the 20 repetitions, I studied the difference between the correlation coefficients as the number of disjointness axioms increases, and plotted a graph with this information.

The graphs in **Figure 5.5** show the result of some of these repetitions. These graphs illustrate that not all disjointness axioms are important for a given dataset. In fact, only for some of the sets of axioms is the correlation coefficient significantly affected, which suggests

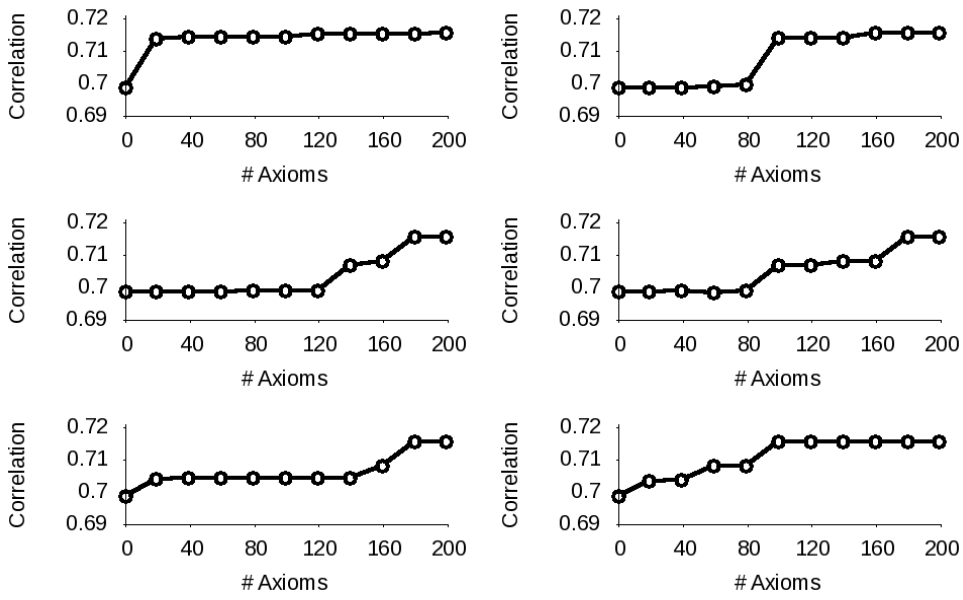


Figure 5.5 – The effect of increasing number of disjointness axioms. These graphs illustrate the increase in correlation coefficient that results from increasing the number of disjointness axioms. In each graph, the abscissa is the number of axioms used by the semantic similarity measure and the ordinate is the correlation coefficient. The correlation coefficient for 0 axioms is always equal to the correlation measured with the classical IC_s, which is 0.69883; the correlation coefficient for the maximum number of axioms corresponds to the value 0.71571, presented in the first validation step. These graphs are representative of the behaviour obtained in all the 20 repetitions.

that those sets contained the axioms that change the logical meaning behind the concepts in the dataset. The graphs present an obvious trend (see **Figure 5.6** for an average of the graphs of all the 20 repetitions) that indicates an increase of the correlation, which, again, indicates that the disjointness axioms improve the correctness of semantic similarity.

3. Effect on other datasets: As the third assessment step, I studied the increase in correlation coefficient on other datasets, since the dataset created for the first step resulted from a random selection process. Following the same selection process, I created 550 more datasets and compared the correlation coefficient as previously explained.

The graph of **Figure 5.7** shows an histogram that represents the difference in the Pearson's correlation coefficient for all these datasets. As is visible in that graph and in **Table 5.1**, the vast majority of the datasets are associated with an increase in the correlation coefficient. In fact, the effect of considering the disjointness axioms for the semantic similarity only impacts negatively 6.2% of the datasets. We observed a mean correlation increase of 0.0149, with a standard deviation for that value of 0.0130. Furthermore, in 72.5% of the datasets, the increase in correlation is significant at a confidence value of 0.05.

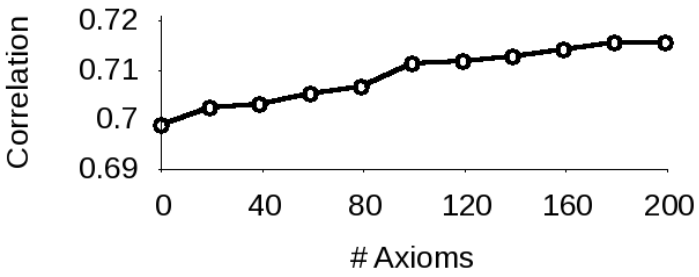


Figure 5.6 – The trend corresponding to all the repetitions. This graph shows the average of all the graphs produced in the 20 repetitions. Although these values do not have any statistical significance in themselves, they clearly show the trend that the more disjointness axioms are considered, the better is the correlation between structural and semantic similarity.

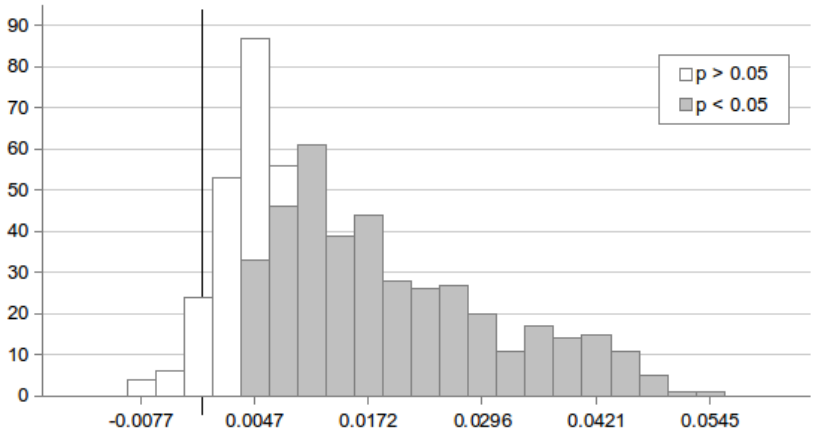


Figure 5.7 – Distribution of the difference in correlation coefficient for random datasets. The majority of the cases show a positive difference. I used Wolfe's t-Test to calculate the p-value associated with the hypothesis that the increase was due to random chance, and marked with a darker shade the amount of datasets for which p-value < 0.05. The vertical line shows the zero of the axis, *i. e.* where the two correlation coefficients are the same.

Table 5.1 – Statistics related to the histogram of Figure 5.7. The last column shows the frequency relative to all the datasets created.

	# datasets	% datasets
Increase in correlation	516	93.8%
p-value < 0.05	399	72.5%

5.1.4 Limitations, future work and other conclusions

This measure has some limitations:

1. The formula of IC_s^{disj} is not robust against ontology development. Sometimes, the addition of a concept to the ontology during development can considerably change the value of shared information content between two concepts (*cf.* Section 4.2.4 "Theoretical validation", where I mention this problem in the context of semantic similarity validation). This issue is mitigated by the fact that the such additions are not common in biomedical ontologies (full details in the paper).
2. The potential for implicit common superclasses is measured using an edge distance, which is a fragile measure in biomedical ontologies [Pes09b] (see Section 3.2 "Edge-based approaches"). It may be possible to explore the semantics of the edges themselves in order to overcome this issue.
3. The measure of information content influences the results obtained with IC_s^{disj}. In this case, IC was calculated with an intrinsic measure of information content; it would be informative to see the effect of changing the IC measure to an extrinsic one.

This validation shows that considering disjointness axioms improves the shared information content measure, with statistical significance. This new approach is able to successfully explore more than just the class-subclass hierarchy of an ontology, relying on a partial subset of the description logic axioms that are included in the ontology to refine the comparison algorithm. To the best of my knowledge, this represents the first attempt to explicitly use description logic expressivity in semantic similarity in the biomedical domain. I demonstrated this hypothesis using a rather naïve approach. More sophisticated approaches include the exploration of the semantics of edges, other types of information content based on external corpus, *etc.*

5.2 Semantic relatedness measure

5.2.1 The idea

According to the definition provided in Section 2.6 "Semantic similarity", semantic similarity uses exclusively the hypernymy relationship of an ontology: the class-subclass hierarchy. In this sense, Heart and Blood are not similar at all. However, in some contexts, hypernymy is not enough to detect that two concepts are related to one another. Anatomy, specifically when used to detect similarity between diseases, is one of these contexts. For example, a heart disease can have implications in blood pressure, and thus a disease annotated with the concept Heart is somewhat related to one annotated with the concept Blood.

I developed a measure that compares two anatomical entities based on their "semantic neighbourhood", an idea that was based on the mental processes that take place in the human

mind when comparing concepts [Qui68; CL75; Tve77] (see Section 3.1 "The art of semantic similarity"). The semantic neighbourhood of a concept is a graph, where each node is a concept and an edge between two nodes is drawn if the two concepts are related in the ontology by means of an existential quantification axiom (see Section 2.3 "Web Ontology Language"). Consider **Figure 5.8**, which shows a snippet of the semantic neighbourhood of Heart as defined in the Foundational Model of Anatomy (FMA). In here, we can see that Heart is related to Aortic valve by the property *has-part*. Many other concepts are related to Heart. The neighbourhood is extended by allowing the neighbours of the neighbours to participate in it as well, recursively. I call the maximal distance between the centre concept and its neighbours the "radius" of the neighbourhood, represented by M and measured in number of edges in the graph. Notice, then, that we can compute the neighbourhood up to any particular radius, and that a wider neighbourhood can convey more information about a particular concept than a narrower one. Using this notion, comparing two concepts is a matter of comparing the two neighbourhoods,

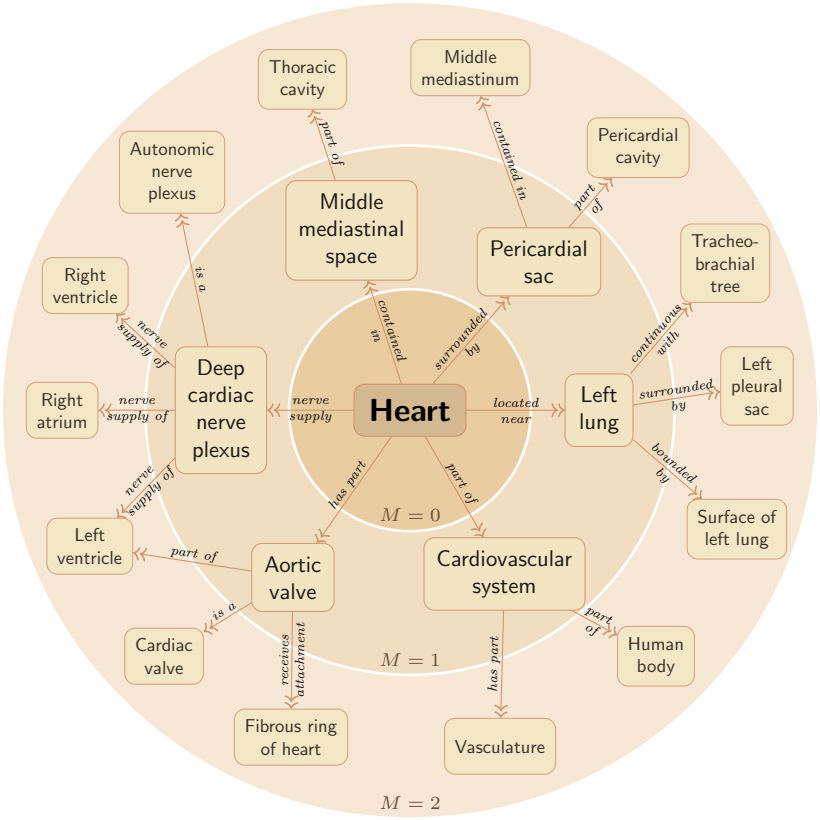


Figure 5.8 – The semantic neighbourhood of the concept Heart. This picture illustrates the semantic neighbourhood of Heart, with a radius $M = 2$. As depicted, Aortic valve is part of the first layer since it directly links to Heart, while Cardiac valve is part of the second layer. Notice that some neighbours can be connected to the centre by more than one distinct path (*e. g.* Left ventricle). The *radius* can be increased, resulting in a larger neighbourhood. Given the proposed weighting mechanism (see eq. 5.12), the closer the concepts are to the centre, the higher is their contribution to the semantic neighbourhood of Heart.

and this was exactly what I proposed [FC11].

Notice that, again, this corresponds to extending the notion of semantic similarity to more logical constructions, in this case the existential quantification. In formal logic, we say that, for each Heart there is a relationship of type *part-of* between that Heart and some Aortic valve:

$$\mathsf{Heart} \sqsubseteq \exists\, \mathit{has\text{-}part}\,.\,\mathsf{Aortic\ valve}$$

Understanding this notation is not essential to appreciate the contributions presented in this document (details and further exploration of description logic symbols can be found, *e. g.* in works by Nardi and Brachman [NB03] and Baader et al. [BHS05]). Simply take into account that this means "Each heart has an aortic valve" (we will encounter this exact notation again in the next chapters).

FMA is one of the best test cases to assess the behaviour of this measure, as it is rich in existential quantifications. In fact, it contains 67 distinct properties (*e. g. part-of*, *surrounded-by*, *continuous-with*, *etc.*), which are used in more than 200,000 axioms.

Finally, the same figure can be used to show the concepts of "weight". The path from Heart to Aortic valve is smaller than the one from Heart to Vasculature, which means that Aortic valve is somehow more related to Heart than Vasculature. I use the notion of "weight" to measure this relative relatedness within a semantic neighbourhood.

5.2.2 The proposed measure

Mathematically, the relatedness between two concepts x and y, is defined as:

$$\mathrm{rel}_{\mathrm{Ferreira}}(x,y) = \frac{\sum_{i \in N_M(x) \cap N_M(y)} \omega(i \to x) \otimes \omega(i \to y)}{\sum_{i \in N_M(x) \cup N_M(y)} \omega(i \to x) \oplus \omega(i \to y)} \tag{5.10}$$

where $N_M(c)$ is the semantic neighbourhood of a concept c calculated to a maximal radius M, and $\omega(i \to c)$ is a weighting function that gives more relevance to the concepts that are closer to c and less relevance to concepts further away.

This formula is relatively complex in syntax, and uses non-standard mathematical operations, and as such needs to be explained bit by bit. In essence, it is a ratio between what is common in the semantic neighbourhoods of x and y, and the total amount of concepts in the two neighbourhoods. This idea is not far from what is used in $\mathrm{sim}_{\mathrm{UI}}$ and $\mathrm{sim}_{\mathrm{GIC}}$ (see eqs. 3.11 and 3.12). Unlike those measures, each concept has now two weights, one for each neighbourhood. To deal with this multiplicity of weights, I used the binary operators of T-norm and T-conorm, represented mathematically by the symbols $\otimes$ and $\oplus$, respectively, which can be applied to two values between 0 and 1. There are several T-norms and T-conorms that could be applied. Mathematically I am interested in the following properties:

$$0 \leq i \otimes j \leq \min(i,j)$$
$$\max(i,j) \leq i \oplus j \leq 1.$$

I chose $x \otimes y = xy$ and $x \oplus y = x + y - xy$ [KMP04].

Consider two semantic neighbourhoods, one for concept A and another for concept B. We want the concepts that belong to both neighbourhoods to increase the overall measure in a way that is related to how important these two concepts are in the two neighbourhoods. For a concept c that belongs to both neighbourhoods, let $w_A = \omega(c \to A)$ and $w_B = \omega(c \to B)$ be the weights of this concept with respect to each of the two neighbourhoods. Consider the following fraction:

$$\frac{w_A \otimes w_B}{w_A \oplus w_B} = \frac{w_A \times w_B}{w_A + w_B - w_A \times w_B}. \tag{5.11}$$

If c is highly relevant in both neighbourhoods (*e. g.* $w_A = 0.8$ and $w_B = 0.9$), both the numerator and the denominator of this fraction have a high value ($\frac{0.72}{0.98}$) and, as such, we observe a small change in the overall $\text{rel}_{\text{Ferreira}}$ (eq. 5.10). If both have a low relevance (*e. g.* $w_A = 0.1$ and $w_B = 0.2$), the numerator will be a low value and the denominator will be a medium-range value ($\frac{0.02}{0.28}$), which contribute to a mild decrease in the overall measure. But if the concept has high relevance in one neighbourhood and low relevance in the other (*e. g.* $w_A = 0.2$ and $w_B = 0.9$), the numerator will be low and the denominator will be high ($\frac{0.18}{0.92}$), which will contribute to a large decrease in the overall measure.

By default, I propose that the weight of a concept with respect to a neighbourhood is computed based on the path that connects that concept to the centre of the neighbourhood. Let p_c be a path connecting concept c to the centre of the neighbourhood. This path is composed of a sequence of properties. For example, in **Figure 5.8**, the path from Heart to Cardiac valve is "*has-part* $\to$ *is-a*". The weight associated to a certain path is the product of the relevance of each of the properties in the path. If more than one path can be traversed from the centre to the concept, then I take the maximum relevance associated with these paths. Formally, let $r(i)$ be the relevance of the property i: then

$$\omega(c \to A) = \max_{p_c} \prod_{i \in p_c} \mathrm{r}(i). \tag{5.12}$$

Finally, the relevance of each property must be predetermined before running this algorithm. I originally proposed using 0.7 as the default relevance, on the basis that it produced the best results from a selection of possible default values (0.6, 0.7 and 0.8). With the passing of time, I came to realise that we could assign each property a relevance that is based on its own information content. Recall the formula used to calculate the information content of a concept based on the ontology alone, proposed by Seco et al. [SVH04] (eq. 3.3). Reusing this formula here, and taking the frequency of a property to be the number of existential quantification axioms in the ontology that use it, we can also define the information content of properties:

$$\mathrm{IC}(i) = 1 - \frac{\log f(i)}{\log N_e} \tag{5.13}$$

where $f(i)$ is the number of existential axioms that use property i and N_e is the number of existential axioms in the ontology.

Furthermore, notice that this measure can both calculate the similarity between concepts and the similarity between annotated entities. In fact, the construction of a semantic neighbourhood can either start on a single concept or on a set of concepts, making this a group-wise relatedness measure.

5.2.3 Validation

The validation strategy followed in this work can be classified as a "Classification prediction". I based this validation on the assumption that anatomical entities implicated in the same disease should be more related than a random pair of anatomical entities. I first created a map between diseases and FMA concepts. To do that, I used the Human Phenotype Ontology (HPO), an ontology that represents abnormalities in human anatomy (such as Abnormality of the eye or Prostate cancer). On the one hand, this ontology is used by its creators to annotate diseases from several disease databases, and associates 6882 HPO concepts with 8013 diseases, with an average and median of 15 and 9 HPO concepts for each disease, respectively. On the other hand, the ontology itself provides semi-formal descriptions of its concepts, with references to FMA concepts. For example, Microtia is described as:

> "Underdevelopment of the external ear (FMA:52781)."

By leveraging on the annotations mentioned above and these FMA references, it is possible to create a dataset of FMA-annotated diseases (see **Figure 5.9**). This dataset can then be used to find pairs of anatomical concept that are implicated in the same disease, which correspond to the positive dataset for this validation. The negative dataset was generated randomly.

The positive and negative datasets were then used to perform Receiving Operating Characteristic (ROC) analysis [Faw06]. This is a common step in classification approaches, summarised as follows:

1. Select a threshold t and create a binary classifier that classifies as positive all the pairs that have $\text{rel}_{\text{Ferreira}} > t$ and as negative the other. For each t, we can determine the true positive rate (TPR — fraction of the related concepts correctly classified as related) and false positive rate (FPR — fraction of the unrelated concepts incorrectly classified as related).
2. The highest threshold results in a TPR of 0 and a FPR of 0 (all pairs are classified as negative); likewise, the lowest possible threshold results in a TPR of 1 and a FPR of 1 (all pairs are classified as positive).
3. Plot the curve defined by the points (FPR, TPR) when the threshold varies from the maximum to the minimum. This is known as the ROC curve. The closer the graph approaches the point $(0, 1)$ (which represents the ideal case where all the positive pairs have a higher relatedness measure than all negative pairs), the better is the measure of relatedness.

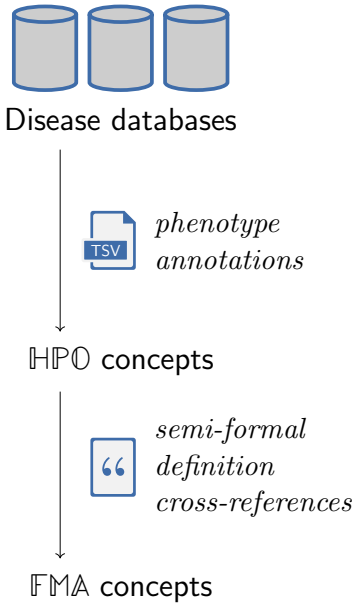


Figure 5.9 – Workflow to find FMA annotations for some diseases. The first step is to find the annotations from diseases to HPO concepts, which are provided as `TSV` files in the HPO website; the second step is to leverage on the cross-references that exist in the textual definitions of the HPO concepts to convert them into FMA concepts. This allows one disease to be associated (*i. e.* annotated) with a set of FMA concepts.

4. Fawcett [Faw04] proposes a way to repeat this experiment a number of times and to produce an average ROC curve from them (see Algorithm 5 of that paper), which I followed by repeating these steps 10 times, each time with a different randomly generated negative dataset.

The ROC curves obtained with this method are presented in **Figure 5.10**. For comparison purposes, I applied the proposed $\text{rel}_{\text{Ferreira}}$ measure, calculated for $M = 3$ and $M = 4$, and I also applied sim_{GIC} to the dataset to study how the behaviour of a similarity measure contrasts with the behaviour of a relatedness measure.

As is evident from this figure, $\text{rel}_{\text{Ferreira}}$ shows a better performance than sim_{GIC}, since high values of TPR are obtained for low values of FPR. The main difference between $\text{rel}_{\text{Ferreira}}$ with $M = 4$ and $M = 3$ is that the former has more resolution power in that it can differentiate between concepts 8 properties apart, whereas in the latter concepts with a path distance greater than 6 have automatically a relatedness value of 0.0. Thus, the measure calculated for many of the negative pairs in the dataset and also some of the positive ones ended up being 0. Given the lower resolution of the measure for $M = 3$, more pairs in this setting have relatedness 0.0, resulting in the straight diagonal line that we see in the figure.

Additionally, sim_{GIC} does not perform as well as the relatedness measure. The distinct form of the graph occurs because a majority of the positive pairs (FMA concepts implicated in the same disease) are not really similar but simply related, just as in the example above: Heart and

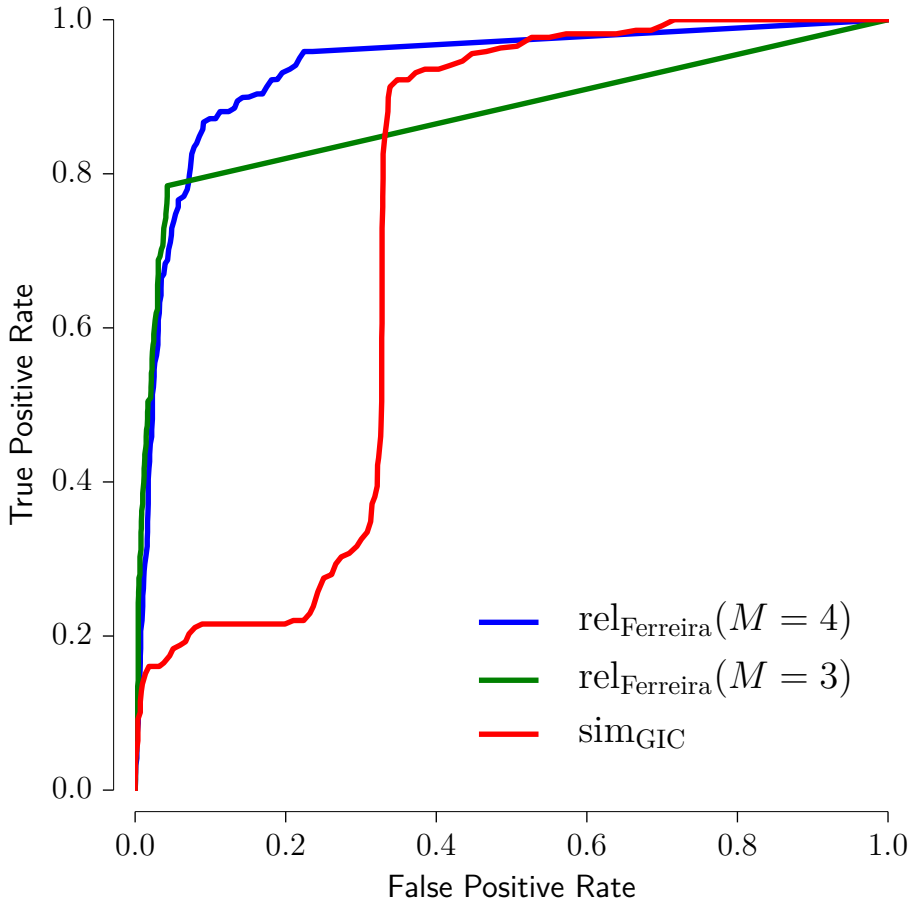


Figure 5.10 – ROC analysis of the results of $\mathrm{rel}_{\mathrm{Ferreira}}$. These ROC curves are the average over 10 runs, each one produced with a different, randomly selected, negative set (*i. e.* randomly selected pairs of FMA concepts).

Aortic valve are relatively frequently implicated in the same disease, and their relatedness is high, as one is part of the other; on the other hand similarity measures are unable to capture this, since one is an organ and the other a valve, two distinct concepts.

This demonstrates the superiority of semantic relatedness measures over semantic similarity, at least when applied to ontologies where there is a vast number of properties, such as FMA, an in contexts where relatedness is more important than mere similarity.

5.2.4 Conclusions and future work

The most important conclusion to take from this work is that measures of relatedness can, in some cases, be more accurate than state-of-the-art measures of similarity, suggesting that relatedness measures do indeed play a role in the biomedical domain, especially when the expressiveness of the relevant ontologies (as measured by the number of properties) is as high as in this case.

The measure that I propose here is based on the concepts of *semantic neighbourhood* and

relevance factors, and can accommodate the needs of particular applications by fine tuning its parameters. For example, by giving different weights to different properties, the measure can give more importance to some neighbours than others. Machine-learning algorithms can be used to tune the weight of each property according to the needs of each application.

The concept of relevant neighbourhood introduced in this work is also a bridge to other methodologies, particularly in allowing the use of ontology mappings to define wider neighbourhoods that draw not only from a specific ontology but from related ontologies as well, as long as a mapping of some sort exists between them. For example, cross-references can be used for this effect. In this context, the measure can incorporate external knowledge, but is not required to do so. For example, the semantic neighbourhood of a disease can include its symptoms and known treatments; to find the neighbourhood of enzymes, it is possible to include the chemical compounds that they transform; *etc.*

Finally, the analysis preformed here on FMA shows that this is a valid method to measure relatedness between biomedical concepts. As we will see later, I have successfully applied this measure to other ontologies and other datasets (see Chapter 7 "Multi-domain semantic measures").

5.3 Summary

This chapter delineates my efforts in incorporating OWL axioms other than hypernymy into semantic measures. The first section deals with disjointness axioms while the second with existential quantification axioms. While this is by no means a comprehensive approach to using OWL formalism in semantic similarity, it paves the way for future experimentation and research in this context. As argued by Couto and Pinto [CP13], increasing the amount of such constructions used in semantic similarity calculations will eventually improve the overall panorama in this field of research and, therefore, the utility of this technique. As such, the results presented here are one of my major contributions to the area of semantic similarity and relatedness.

Even though these results do not directly support the idea of multiple-ontology semantic similarity, they were invaluable to the whole corpus of research that I was committed to achieve. In fact, they represent real scientific advances. Additionally, although $\text{rel}_{\text{Ferreira}}$ was not originally validated in a multi-domain context, it can be easily converted into a measure that is able to tackle that problem, as we will see in Chapter 7 "Multi-domain semantic measures", by allowing the semantic neighbourhood of concepts to be generated based on several ontologies.

Interlude

At this point in the document, it can perhaps benefit the reader to make a small summary of what has been discussed so far and how I continue to delineate my scientific contributions.

Multi-domain semantic similarity is useful to compare resources whose description spans several domains of knowledge, and biomedical informatics is rich in such resources: *e.g.* epidemiological surges can be described using diseases, symptoms, pharmaceutical drugs, geographical locations, *etc.* Comparison of these resources is important to enable searching capabilities on the multidisciplinary datasets, while allowing a certain kind of "fuzziness" on this search (resources need not fully satisfy a user query but can, instead, be similar to it).

Semantic similarity has been traditionally developed for single ontologies. As of June 2015, some published works deal with the multi-ontology problem, but all of them use multiple ontologies of the *same domain* of knowledge, in an attempt to complement the knowledge in one ontology with the knowledge in another. Multi-*domain* semantic similarity, in opposition, is important to compare multi-domain resources, annotated with concepts not only form distinct ontologies but from different fields of knowledge. No published literature deals with this problem, as far as I know.

As such, multi-domain semantic similarity measures need to be developed and validated. This is the research focus which this document reports. On the one hand, we can use single-ontology measures to compare concepts from one entity with "compatible" concepts from the second entity, thus obtaining a set of similarity values that can be mathematically aggregated into a single similarity value (aggregative approach); on the other hand, we can integrate all the relevant ontologies in a single knowledge-base and use existing measures directly on top of it (integrative approach).

This multi-faceted task will be described in the next three chapters. Chapter 6 "Multi-domain data" describes three multidisciplinary datasets collected to serve as test cases for multi-domain semantic similarity. The multi-domain measures (aggregative and integrative approaches) are presented in Chapter 7 "Multi-domain semantic measures", along with their associated results, stemming from its application over the three datasets. Finally, Chapter 8 "Semantic similarity software suite" will examine the technical details of semantic similarity, with particular focus on an open-source software suite that I developed to assist in calculating semantic similarity using OWL ontologies.

Chapter 6

Multi-domain data

Although there is a need for multi-domain semantic similarity measures in the generality of the semantic web community, specifically within the scientific community, there is still a lack of substantial data this technique can be applied to. This, it can be argued, seems a contradictory state of affairs. Either there is data and as such the techniques to analyse them are needed, or there is a lack of data and the techniques are superfluous.

The truth is that semantic similarity is not a *pressing* need for state-of-the-art semantic web practices, nor is it fundamental to *current* scientific progress. In generic areas of the semantic web (outside the biomedical scope), ontologies are not even highly used, and knowledge is not *well* represented, by which I mean *i*) that ontologies are not consistent, either internally or with the external world; *ii*) that the represented knowledge is severely incomplete; or *iii*) that there is no way to integrate that knowledge with other ontologies, since the principles of interoperability are overlooked or neglected. For example, `dbpedia.org` [Biz09] is a collection of information spanning most domains of knowledge (based on the structured information of Wikipedia), but it is particularly rich in instance-level properties, not in ontological knowledge: while it contains the information that "Lisbon" *has-timezone* "Western European Time", there is no representation of geospatial knowledge, which would, in particular, contain the axiom that *has-timezone* is a property that can be applied to instances of Place and whose value must be a Timezone. Nevertheless, there is a small number of ontological information expressed in OWL, such as "Capital *is-a* City".

In contrast, while some areas of research, particularly in biomedical domains, are developing an increasing number of ontologies, there is still a lack of data annotated with them. One notable exception is the Gene Ontology, which is extensively used to annotate proteins and the results of genomic experiments (see Section 2.5 "Semantic annotation" and Chapter 4 "Validation strategies"). This exception suggests that the lack of data does not correspond to a fundamental characteristic of scientific knowledge; instead, I argue that there is no motivation to annotate in the first place because there are not many tools able to explore the data. Once these tools start to appear, more data will be emerge. In light of this, one of the tasks I executed was multi-domain semantic annotation acquisition. These data allow the application of similarity measures on real datasets.

In this section, I present three datasets that were collected for exploiting semantic similarity, in three different areas of research: epidemiology, metabolism, and computational modelling of biological processes. For a list of the ontologies used by these datasets, see Appendix A "List of ontologies".

6.1 Epidemiology Dataset

Epidemiology is inherently a multidisciplinary subject, relying on areas of knowledge as diverse as medicine, biology, statistics, sociology and geography [Por08]. Even under the scope of medicine and biology, epidemiology deals with chemical compounds, diseases, symptoms, environmental conditions, methods of transmission, vaccines, *etc.*

Given this multitude of domains, processing, storing and preserving epidemiological data is not straightforward. To explore ways of managing this type of data, a consortium of several partners established the Epiwork project, aimed at developing the appropriate framework of tools and knowledge to design epidemic forecast infrastructures, including an epidemiology data repository that was developed by the LaSIGE partner (see Section B.1 "Semantic web in the Epidemic Marketplace" for a summary of my contributions to the Epiwork project).

One of the most important functions of a data repository is the ability to search within its resources. A search box that can be used to convert a user query into a list of results is essential for the widespread adoption of the repository. There are at least two possible ways to implement this feature: *i*) allow free text searches that try to map the words in the query to the words in the content of each resource; or *ii*) annotate the resources of the repository with metadata that reflect its content and use the query to search within these metadata. While the first way maps roughly to how web search engines work today, the second way is much more aligned with the idea behind the semantic web, with all its advantages (see Sections 2.5 "Semantic annotation" and 2.4 "Semantic web") and, as such, the team behind the Epidemic Marketplace decided to provide a way for users and curators to annotate their resources.

Given the multidisciplinary nature of epidemiology and the resources contained in the Epidemic Marketplace, a multi-domain semantic similarity measure would be an asset to assist the search functionality, which would require a means to compare resources based not on one domain in particular (*e. g.* diseases), but on all the domains of annotation.

Consider a user searching for resources about "infectious diseases". Using the Human Disease Ontology (𝔻𝕆𝕀𝔻), specifically its class-subclass hierarchy, the search functionality can successfully retrieve all resources with data on Flu, AIDS and other infectious diseases. However, it can happen that the user queries for very specific resources (for example by over-specifying the disease, the symptom or the location of an outbreak), resulting in an empty list of retrieved resources. In such cases, it is possible to find some results that *almost* satisfy the query by comparing the repository resources with the query. For example, the query "2009 European infectious disease outbreaks that manifest through coughing" could be satisfied with a resource about

a 2010 European infectious disease outbreak that manifested through sneezing. Additionally, when the query does return some results, semantic similarity can be used to sort those results according to how related they are to the original query.

On the other hand, semantic similarity provides a mechanism to implement a "Related resources" section in the repository. Users looking at the contents of a particular resource are usually interested not on a single resource but on a collection of them, all of which are related. For example, a user looking at a resource that contains data about flu-like diseases is probably also interested in resources about other infectious pulmonary diseases. Having a section of the web page dedicated to these related resources removes the need for the user to make complex and sometimes unintuitive queries to the search feature.

One of my contributions to this project was the development of a semantic metadata model and a Network of Epidemiology-Related Ontologies (NERO). Both are used to assist the annotation of epidemiological resources: the metadata model describes the type of information that a resource needs and NERO provides concepts for the annotation (see Section B.1 "Semantic web in the Epidemic Marketplace").

During the Epiwork project, it was possible to annotate a set of 228 resources, each one containing a reference to an open-access paper from an epidemiology journal (the annotation process itself was conducted by someone else in the project, not me). Each of these resources is annotated with a set of concepts from NERO according to the semantic metadata. Hence, this dataset contains metadata on domains such as "environment", "diseases", "symptoms", "modes of transmission", "demography" *etc.*

By leveraging on NERO to represent the concepts that each paper refers to, the papers become enriched with semantic information and can, therefore, be used in semantic analysis. While there is no explicit quantitative assessment of similarity or relatedness between these resources, and as such there is still no gold-standard that can be explored in the development of multi-domain semantic similarity, doing so is not beyond the realm of possibilities, since the resources are already annotated. As such, I gathered this dataset to run semantic similarity on it, as we will see later in Chapter 7 "Multi-domain semantic measures".

A summary of the relevant annotations for these resources is given in **Table 6.1**. In this table, I make use of three statistics that depict the annotation panorama of these resources with respect to a single domain:

Coverage This is the fraction of resources that have at least one annotation in the domain.

Volume This is the average number of concepts from this domain in the resources. It is calculated with respect to the resources that have at least one annotation, which means that the minimum value is 1.0.

Diversity This is the number of distinct concepts from the domain that are used throughout all the dataset.

Table 6.1 – Summary of the annotation in the epidemiology dataset. This dataset corresponds to 228 epidemiology resources extracted from the Epidemic Marketplace. The first two columns describe the domains and ontologies used to annotate the resources. The rest of the columns provide statistics for each domain. A description of each statistic is given in the text.

Domain	Ontology	Coverage	Volume	Diversity
Chemistry	CHEBI	0.9%	1.00	1
Diseases	DOID	59.6%	1.76	70
Environment	ENVO	21.1%	1.00	9
Phenotypes	PATO	0.9%	1.00	1
Symptoms	SYMP	46.1%	3.55	79
Transmission	TRANS	42.5%	1.00	9
Vaccines	VO	20.6%	1.06	16
General	NCIt	100.0%	4.13	157
General	MeSH	83.8%	2.24	131

6.2 Metabolic Pathways Dataset

Another multidisciplinary area in the field of biomedical informatics is metabolism. This field studies the chemical reactions that take place in a living organism and which are the basis of biology and life in general. For example, it studies how the energy of sunlight is used by plants and other organisms to convert water and carbon dioxide into oxygen and glucose, a process known as photosynthesis. A full description of such a process is called a *metabolic pathway*, which is often depicted as a graph showing the intervening molecules.

The process described by a metabolic pathway is usually performed inside a cell, or within its immediate surroundings, with the assistance of enzymes (proteins that accelerate the chemical reactions), and it has certain chemical inputs and outputs. Metabolic pathways encompass several smaller steps (the individual chemical reactions) and several intermediary molecules, such as the metabolites (the molecules that are transformed), the enzymes, and other regulatory proteins that supervise the whole process based on cellular conditions (such as the amount of oxygen within the cell, the amount of sunlight, *etc.*). **Figure 6.1** contains a simple example of a metabolic pathway involving several metabolites and enzymes.

While these chemical reactions usually occur within living organisms as a continuum, *i. e.* there is no naturally defined boundary between metabolic pathways, dividing the high amount of distinct chemical reactions into manageable groups simplifies the study of metabolism. For example, the glucose produced by photosynthesis, in plants, is converted into other molecules, but even though the two processes happen simultaneously, each is described by a distinct pathway.

Knowing the components of a metabolic pathway and how they interact with one another is extremely helpful: *i*) a metabolic pathway is a means of effective communication regarding

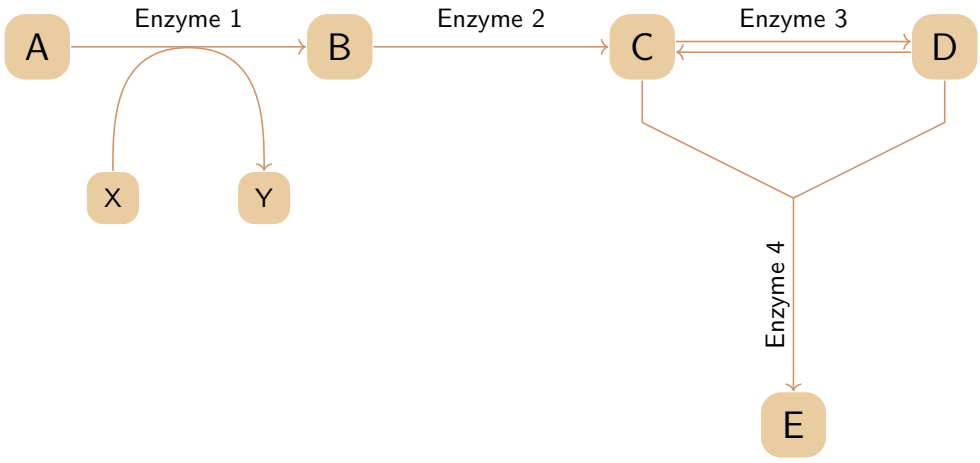


Figure 6.1 – Example metabolic pathway. This figure illustrates a hypothetical metabolic pathway. Chemical reactions are catalysed by specialised proteins known as enzymes, which accelerate the rate at which metabolites are converted. The image depicts that some reactions require extra metabolites, and other produce extra chemical compounds. The pointed arrows represent the "flow of matter", *i. e.* the fact that one metabolite is converted to another one, thus implicitly describing the input and output of the pathway. In this case, the inputs are the metabolites A and X, and the outputs are the metabolites E and Y.

the metabolism of various organisms [Pap03]; *ii*) it enables the mathematical analysis of a complete metabolic system, since they correspond to precise mathematical descriptions of cellular properties [Pap03]; and *iii*) this information can provide insight into how organisms respond to failures in the pathway (such as the absence of an enzyme) [Bau11]. Furthermore, inside cells, many different processes occur at the same time, with thousands of molecules being converted concurrently. As such, to quickly estimate the effect of changing something in a pathway, assistance from computerised system is required.

What makes metabolic pathways multidisciplinary is the fact that, to fully describe them, we need not just the metabolites that are converted and the molecular functions that are carried out, but also the drugs that interfere with the pathway and the diseases caused by pathway defects or other malfunctions. The Kyoto Encyclopedia of Genes and Genome (KEGG — http://www.genome.jp/kegg/kegg2.html) is a collection of databases on biological systems. One of these databases is KEGG PATHWAY, which categorises 269 pathways into a hierarchy, and annotates each pathway with: *i*) chemical compounds, *ii*) enzymes, *iii*) drugs that affect the pathway, and *iv*) diseases that are associated with the pathway. Annotation is done with concepts from other KEGG databases. The concepts linked to in these annotations are not from the reference ontologies mentioned in Appendix A "List of ontologies", but conversion to GO and CHEBI can be performed, using KEGG's own internal tables, which map compound and drug concepts to CHEBI, and genes to UniProt identifiers (which can be used to find GO annotations for the genes). Diseases, however, have no link to reference ontologies. As such, an additional step was executed to convert KEGG diseases into Human Disease Ontology (DOID) identifiers.

Table 6.2 – Summary of the annotation in the metabolic pathways dataset. This dataset corresponds to 269 pathways extracted from KEGG. Note that a domain in this dataset does not precisely correspond to an ontology, as both drug and metabolite annotations use concepts from CHEBI. The first two columns describe the domains and ontologies used to annotate the resources. The rest of the columns provide statistics for each domain.

Domain	Ontology	Coverage	Volume	Diversity
Diseases	DOID	79.2%	9.08	756
Drugs	CHEBI	42.4%	30.30	1381
GO terms	GO	100.0%	32.45	3210
Metabolites	CHEBI	79.9%	21.37	2628

Table 6.2 shows a summary of the annotations for these pathways, using the same statistics presented for the previous multi-domain dataset. Notice that in this dataset, there is not a one-to-one correspondence between domains and ontologies, since both drugs and metabolites are represented as CHEBI concepts. I decided to keep the two domains separate since they encode different information about the pathways.

Like in the case of epidemiology resources, a database of pathways needs a multi-domain semantic similarity measure to be able to properly make use of all the information available about each pathway in order to answer user requests with the most relevant resources. Additionally, semantic similarity in metabolic pathways can also be used to *i*) reconstruct phylogenetic trees depicting the common metabolic history of a group of organisms [HS03], and *ii*) find suitable model organisms in the study of diseases related to a certain metabolic condition [FS99]. Single-ontology semantic similarity has been researched in this domain by Clemente et al. [CSV05], which used similarity of proteins based on their GO annotations, and Grego et al. [Gre10], which used instead semantic similarity between the metabolites of the pathways based on CHEBI.

By gathering metabolic pathway information in this way, I effectively created a dataset of pathways annotated with CHEBI, GO and DOID concepts, which was further used in semantic similarity studies (see Chapter 7 "Multi-domain semantic measures").

6.3 Biochemical Models Dataset

A third area where multi-domain semantic similarity can be of service corresponds to models of biological systems, or biomodels, for short. This type of information is similar to the one of the previous dataset: like a metabolic pathway, a biomodel is a description of a chemical process that happens inside the cell, or within its immediate surroundings. In contrast, however, a biomodel is more computationally oriented. It represents the metabolites and enzymes involved in the reactions but also the cellular components and anatomical location where they occur, including their volume, and the equations that describe the reaction velocity with respect to the concentration of the metabolites and enzymes.

Table 6.3 – Summary of the annotation in the biomodels database. This dataset corresponds to the 282 distinct biomodels extracted from the BioModels website. The first column describes the ontologies used to annotate the resources. The rest of the columns provide statistics for each ontology.

Ontology	Coverage	Volume	Diversity
CHEBI	55.0%	6.99	261
FMA	3.9%	1.18	11
GO	90.8%	55.43	3314
PATO	95.4%	1.06	5

Like in the previous two scenarios, multi-domain semantic similarity is useful here as well, primarily to enable searching capabilities within a database of biomodels. Other applications include *i*) clustering the biomodels according to similarity in order to find common patterns in different organisms, which can be used to transfer knowledge from one organism to another; and *ii*) using semantic similarity to assist the act of annotating the models, by analysing similar biomodels and generating new annotation suggestions to increase the accuracy of the annotations given by an author [SKL12].

The EBI Biomodels Database (http://www.ebi.ac.uk/biomodels-main/) contains formal descriptions of mathematical models of biochemical systems [Li10a; Jut15]. These biomodels are annotated with concepts from the relevant domains, including chemical compounds, enzymes, biological processes and anatomical entities.

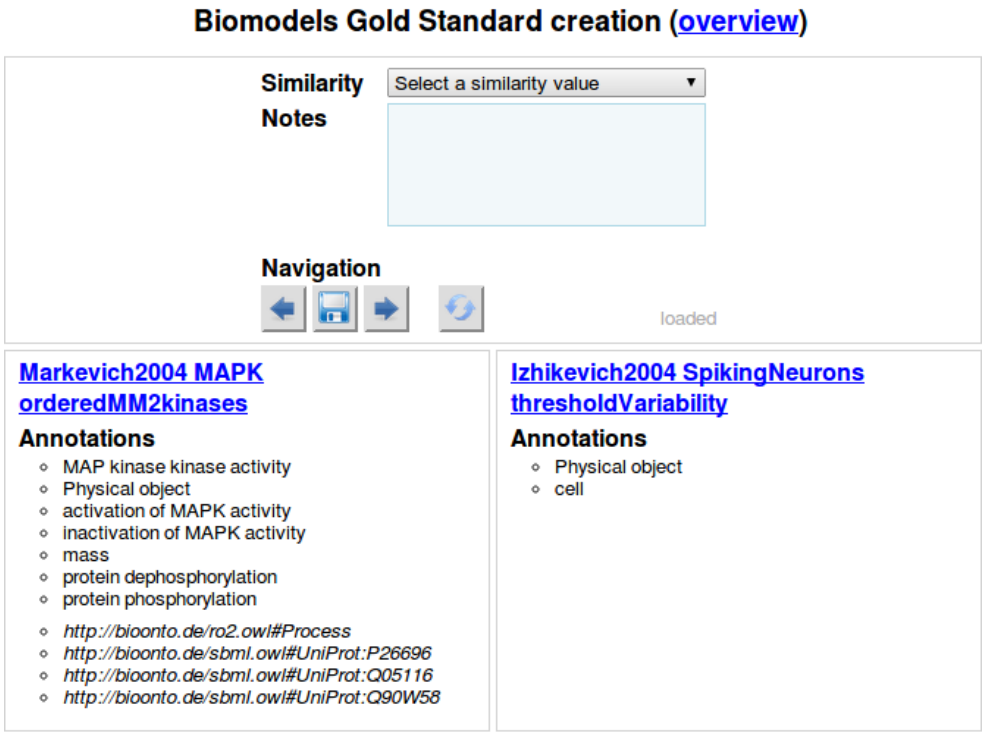
The models in this dataset are annotated with the reactions that they represent, the chemical compounds involved in the reactions (both metabolites and enzymes) and the cellular components where the reactions occur. This information is frequently (but not always) accompanied with links to ontologies such as GO and CHEBI. For example, reactions are linked to concepts from GO; chemical compounds are linked to CHEBI concepts, KEGG COMPOUND terms, and InterPro and UniProt identifiers; protein complexes that participate in the reaction and cellular components are linked to the Cellular Component branch of GO (which represents complexes as well as membrane-delimited components); cell components are also linked to FMA concepts; and physical quantities, like mass and electric charge, are linked to PATO concepts.

Given the complexity of these annotations, I decided to make some conversions and ignore some annotations. For example, I ignored KEGG PROTEIN terms (there are only 3 in the whole set of annotations). KEGG COMPOUND terms were converted to CHEBI concepts when possible (compounds without a correspondence were ignored as well), and UniProt and InterPro identifiers were converted into the GO annotations for those proteins. This resulted in each model having annotations to CHEBI, FMA, GO and PATO. A summary of the annotations for the biomodels is given in **Table 6.3**.

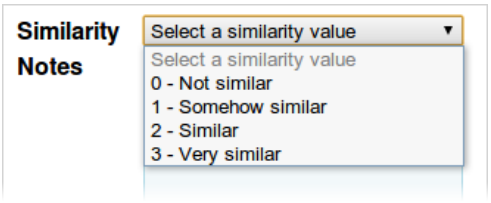
I extracted from this dataset 250 pairs of biomodels (for a total of 282 distinct biomodels), with the aim of having a Systems Biology expert assess the degree of similarity between each

pair. The data in this *gold-standard* corresponds to the models and their annotations from the corresponding ontologies. To ensure a good coverage of all similarity values, 100 pairs were generated randomly and the other 150 were generated based on a preliminary semantic similarity calculation, in order to have a balanced distribution of similarity values. To this effect, I first calculated semantic similarity on all the biomodel pairs and divided the pairs into three categories: one for similarity values below 0.33, another for values between 0.34 and 0.67, and another for values higher than 0.67. This ensured that pairs covering the full range of similarity were included in the gold-standard. The 100 random pairs were generated to cover the possibility that the preliminary similarity values were not significant.

The 250 pairs were classified by the expert as "not similar", "somehow similar", "similar", and "very similar". Expert assessment was conducted based on a web-tool that I designed for that effect. **Figure 6.2** displays some screen-shots of the tool.



(a) The main page of the similarity assessment tool



(b) A detail of the similarity selection section

Figure 6.2 – The EBI Biomodels similarity assessment tool. (a) Each model is accompanied with the set of ontology concepts that annotate it. In the bottom of this panel, there are control buttons to navigate through the 250 pairs and a "Notes" section to allow the expert to make their own notes, for future reference. (b) The similarity value can be chosen between 0 and 3.

Chapter 7

Multi-domain semantic measures

Given the multidisciplinarity of biomedical resources, it is necessary to implement measures of similarity that can handle all the relevant domains. For instance, to compare metabolic pathways one can use protein similarity [CSV05] or chemical similarity [Gre10], but we can also conceive a scenario where both sources of knowledge are important and, therefore, where comparison needs to explore the two domains. In theory, this should provide a more accurate insight into what the pathways represent in the real world and, ultimately, contribute to a better similarity measure.

In this chapter, I propose the hypothesis that multi-domain semantic similarity has some advantages compared to classical single-ontology measures when dealing with multidisciplinary resources; in particular, I demonstrate this hypothesis based on the accuracy of both techniques in the three different datasets presented in Chapter 6 "Multi-domain data". By doing so, I also show that multi-domain measures are necessary for the advancement of science and that, as time passes and the amount of resources annotated with concepts from multiple ontologies increases, the demand for such measures will also increase.

Given the success of single-ontology semantic similarity measures in the past, instead of a new measure of semantic similarity built from scratch, I propose two mechanisms that *lift* single-ontology measures into their multi-domain counterparts: the *aggregative approach* compares each of the domains of relevance independently and then aggregates the several similarity values into a final score; and the *integrative approach* integrates all the ontologies under the same common root and then applies single-ontology measures on it.

7.1 The two multi-domain approaches

Multidisciplinary entities are commonly annotated in domains that are on a one-to-one correspondence with a set of ontologies, since biomedical ontologies tend to represent a specific domain of knowledge, per the guidelines of the biomedical informatics community (see Section 2.2 "Ontologies" and Appendix A "List of ontologies"). For example, concepts used to annotate epidemiology resources include diseases from DOID, symptoms from SYMP, vaccines from VO, *etc.* Sometimes, the same ontology corresponds to more than one domain (*e. g.* in metabolic pathways, annotations to CHEBI are used both for metabolites and drugs, and GO annotations for molecular functions, biological processes and cellular components); the reverse is not as common (more than one ontology annotating for the same domain).

My work assumes that dividing the annotations in domains can be done in a straightforward way (*cf.* the datasets described in Chapter 6 "Multi-domain data", where such division is presented in those chapter's tables). It also assumes the existence of a group-wise single-ontology semantic similarity measures which can compare a set of concepts with another set of concepts, examples of which include:

- concept-wise similarity measures, such as $\mathrm{sim}_{\mathrm{Resnik}}$ (eq. 3.4) and $\mathrm{sim}_{\mathrm{Lin}}$ (eq. 3.7), which can be made group-wise with the use of an aggregation technique, as described in Section 3.5 "Comparing annotated entities"; and
- measures that are inherently group-wise, like $\mathrm{sim}_{\mathrm{UI}}$ (eq. 3.11), $\mathrm{sim}_{\mathrm{GIC}}$ (eq. 3.12) and $\mathrm{rel}_{\mathrm{Ferreira}}$ (eq. 5.10).

The following subsections describe the two approaches that lift single-ontology measures into multi-domain measures.

7.1.1 Aggregative approach

This approach treats each domain of annotation independently. For each annotation domain, the concepts of that domain used to annotate the first entity are compared to the concepts of that domain used to annotate the second entity, using the group-wise single-ontology measure. This produces a collection of similarity values, one for each domain, which must then be aggregated with the use of a function such as the maximum, the minimum or the average. See **Figure 7.1** for a graphic illustration of this process.

While the aggregation technique can be one of several different options, I show here only the results of two of them: the raw average and the weighted average. Let e and e' be two multi-domain annotated entities being compared, e_d and e'_d be the set of concepts annotating the entities e and e' in the domain d, D be the set of all domains annotating the two entities,

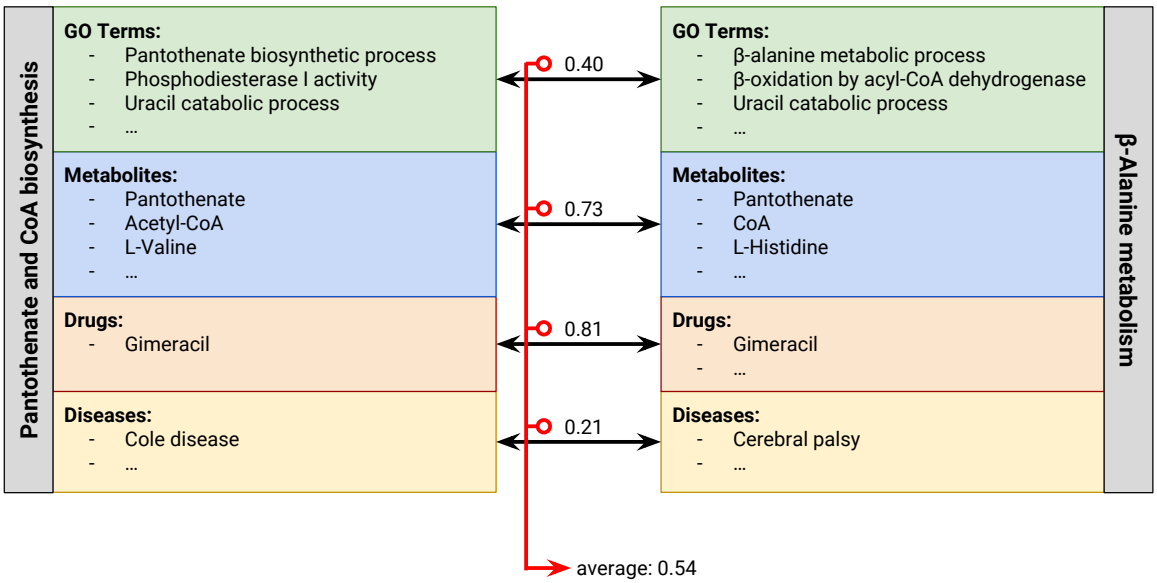


Figure 7.1 – The aggregative approach. This image illustrates how this mechanism works using two metabolic pathways as example. Each domain is identified by colour. The concepts in each domain in the first entity are compared to the concepts in the same domain in the second entity, and then the values are aggregated; in this example, the aggregation mechanism was the raw average.

and σ be the single-ontology semantic similarity measure being lifted. I define

$$\text{sim}_{\text{Aggr}_{\text{raw}}}(e, e') = \frac{1}{N} \cdot \sum_{d \in D} \sigma(e_d, e'_d) \tag{7.1}$$

$$\text{sim}_{\text{Aggr}_{\text{weighted}}}(e, e') = \frac{1}{\sum_{d \in D} w_d} \cdot \sum_{d \in D} w_d \cdot \sigma(e_d, e'_d) \tag{7.2}$$

where w_d is the weight associated with the domain d. I followed an approach that weights each domain by the amount of annotations that it provides to the entities being compared; as such, domains that are more represented in an entity contribute with a higher weight to the final similarity value:

$$w_d = \left| e_d \cup e'_d \right| . \tag{7.3}$$

For example, if a domain contributes to the annotations of e and e' only with one concept (either only to one of the entities or to both), the weight of this domain will be 1. Domains that contribute with a higher number of concepts have a higher weight on the overall similarity value.

This method has the advantage that it can directly use existing measures to compute similarity and works irrespective of the degree of interoperability between the various ontologies used to annotate the entities.

7.1.2 Integrative approach

While one of advantage of the previous approach is the possibility to be applied to non-interoperable ontologies, biomedical ontologies are *generally* interoperable in at least two ways:

- the use of the Basic Formal Ontology (BFO) as an upper ontology [GSG04], which means that concepts form different ontologies have the potential to share common superclasses (even if only general and abstract ones); and
- the use of cross-references between ontologies (*e. g.* the link between a disease and the anatomical entities that it affects), which most ontologies in this field try to satisfy. This stems from the reuse of concepts from different ontologies (*e. g.* VO reuses the concepts Chemical entity from CHEBI and Protein complex from GO), which enables ontologies to refer to concepts outside their domain but at the same time relevant for describing the knowledge of that domain.

In fact, by separating the various domains into independent groups, we lose information that could also be used to compute similarity. First, if only one of the entities is annotated in one domain, (*e. g.* FMA), this domain is effectively ignored, and no amount of annotation can change that. Second, inter-domain relationships between concepts in different ontologies are also ignored. For example, an annotation to the concept Deafness cannot be correlated with an annotation to Ear or even Hearing, since those concepts are all part of different domains (they are, respectively, represented in, DOID, FMA and NCIt).

One way to avoid the pitfalls of separating the multiple domains in isolate computations is to merge all the relevant ontologies in a single multi-domain virtual ontology and then to use the single-ontology measure directly on top of this virtual ontology. This is the second approach, where all concepts of one entity are compared to all the concepts of the other (see **Figure 7.2**). Measures of this type assume, therefore, that there is interoperability between the ontologies being used. This is essential in two accounts:

- in relatedness measures, it is vital, for example, that the molecular function ATP binding is explicitly related to the chemical compound ATP, since this relationship allows the relatedness measure to compute a high value for this pair of concepts;
- likewise, although the molecular functions Ethanol degradation and Cellular response to ethanol, both part of GO, are not fundamentally similar (one is the process by which the body converts ethanol to other smaller molecules, the other is the process that cells undergo in the presence of ethanol, and their most informative common superclass is the abstract concept Physiological process), both contribute to the metabolism of the same compound, and are similar in the sense that the two processes are related to Ethanol, a concept that is represented in a different ontology. Exploring this relationship can also increase accuracy of similarity.

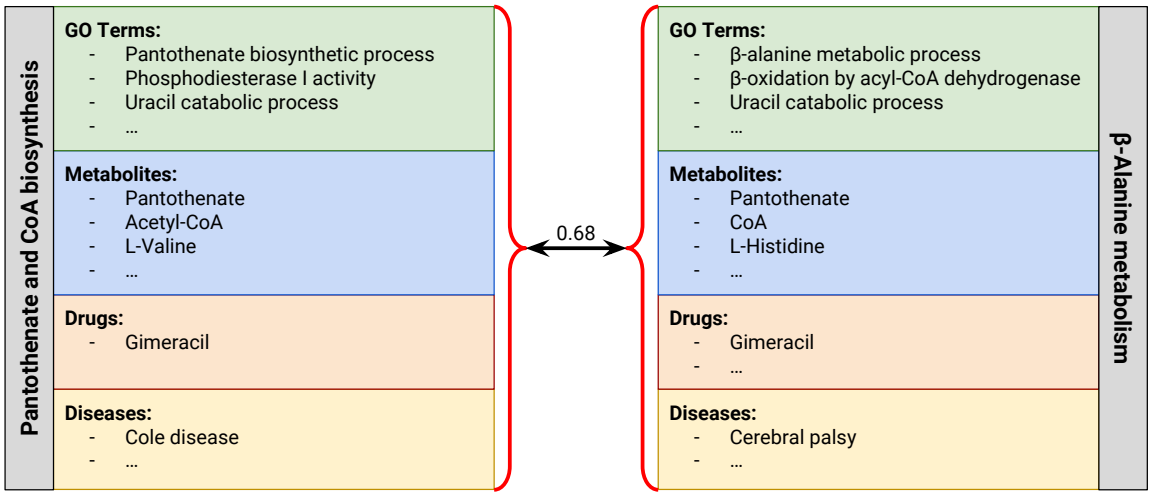


Figure 7.2 – The integrative approach. All the ontologies are aggregated under the same root, which means that single-ontology measures can be directly applied to compare concepts from different domains. In this case, only one value is obtained, which corresponds to the semantic similarity value.

To achieve this, all the ontologies that are used must be in some way merged into a single knowledge representation artefact (a single ontology). Merging ontologies is a process that is mainly studied by Ontology Matching [SE05; ES07], and consists in automatically or semi-automatically finding the concepts that are equivalent in two ontologies. In the semantic web context (see Section 2.4 "Semantic web"), ontology matching has a critical role, since it helps integrate heterogeneous entities, usually created by different research groups but with a certain overlap in their domains. However, there is a large effort to make biomedical ontologies interoperable, in the sense that *i*) concepts are reused among ontologies to refer to the same real-world idea, *ii*) there is a common upper level ontology, which represents the most abstract concepts and provides a common ontological background that enables an objective classification of concepts, and *iii*) ontologies are orthogonal, thus each is responsible for representing exactly one domain of knowledge. As such, ontology matching between reference ontologies in the biomedical domain is not expected to produce a high number of matches. In fact, since each ontology represents a different domain, there are, theoretically, no "semantically equivalent" concepts in any pair of reference biomedical ontologies (in practice, this is not exactly true, as some domains overlap — *e. g.* both GO and FMA contain the concept of Cell, represented with different identifiers — but the amount of overlap is extremely small, it is decreasing as time goes by, and it is almost always relative to abstract rather than specific concepts).

Therefore, in practice, biomedical ontologies that need concepts from other domains "import" those concepts from other ontologies. For example, many biological process concepts in GO

represent biochemical reactions, and these concepts are appropriately linked to the CHEBI concepts that represent the molecules that are transformed in these reactions: *e. g.* the concepts Ethanol degradation and Cellular response to ethanol mentioned above are linked to the concept Ethanol represented in CHEBI. These facts are asserted using existential quantification axioms (see Section 5.2 "Semantic relatedness measure"):

$$\text{GO:“Ethanol degradation”} \sqsubseteq \exists\, \textit{has-input}\,.\,\text{CHEBI:“Ethanol”},$$
$$\text{GO:“Cellular response to ethanol”} \sqsubseteq \exists\, \textit{has-input}\,.\,\text{CHEBI:“Ethanol”}.$$

Rather than ontology matching, it is these inter-domain cross-references that can potentially increase the accuracy of semantic similarity measures. With them, it becomes possible to find a specific rather than abstract connection between the GO concepts Ethanol degradation and Cellular response to ethanol, which was otherwise absent from the GO ontology. Measures such as $\text{rel}_{\text{Ferreira}}$ are capable of exploring the inter-domain cross-references and use them to compare ontology concepts (see Section 5.2 "Semantic relatedness measure") and, by extension, annotated entities.

Unfortunately, the current state of cross-linking in biomedical ontologies is largely underdeveloped, despite it being a recommended practice by the OBO Foundry. The group responsible for developing and maintaining GO intends to provide cross-references to appropriate concepts from other ontologies. Examples already deployed are the ones given in the previous paragraph, which link GO and CHEBI. Planned links include cross-references to anatomical locations and species names [Mun11]. Furthermore, HPO represents human phenotype abnormalities, and links them to anatomical concept from FMA. For example, Microtia is defined as the "Underdevelopment of the 'external ear' (FMA:52781)".

Incidentally, OWL is well suited to deal with this type of cross-reference. If two ontologies refer to the same identifier, when the ontologies are used together, that identifier will refer to the same concept; as long as all the necessary ontologies are loaded, these type of inter-domain axioms can be defined in one ontology using an identifier from another ontology, which makes this multidisciplinary fact directly available out of the box. This is because, in the semantic web, identifiers are universal, and always identify the same concept. For example, there is an OWL file representing GO that contains the axioms exemplified above, which relate molecular functions with chemical compounds. This file does not have any information about the CHEBI concepts other than the identifier. Therefore, an automatic system that needs more information on those CHEBI concepts, *e. g.* its superclasses, name and synonyms, must load CHEBI to find it.

7.2 Results

I have applied semantic similarity to the multi-domain datasets described in Chapter 6 "Multi-domain data". As these are intrinsically multidisciplinary datasets, they provide an appropriate

and highly relevant testbed to try the two approaches described above. I refer the reader to **Table 6.1** (page 84), **Table 6.2** (page 86) and **Table 6.3** (page 87), describing the amount of annotations on these datasets, which will, therefore, be pertinent to the analysis presented here.

Except where otherwise stated, all the results presented in this section were obtained using $\text{sim}_{\text{Resnik}}$ (eq. 3.4) as the group-wise single-ontology semantic similarity measure. Since this is a concept-wise measure, I used it to create a similarity matrix, where each value is the similarity between one of the annotations in the first entity and one of the annotations in the second entity, and then use a Best Match Average (BMA) approach to convert this matrix into a single similarity value. Other group-wise single-ontology measures have been used, in particular $\text{rel}_{\text{Ferreira}}$, with results equivalent to the ones presented. As such, these results are not shown, except where relevant. In particular, although some measures are better suited to tackle some problems than other measures (for example, a measure that uses disjointness axioms is better suited to deal with problems related to ℂℍ𝔼𝔹𝕀 concepts, as detailed in Section 5.1 "Disjointness axioms in semantic similarity"), the increase in performance observed in multi-domain semantic similarity over single-ontology semantic similarity is largely independent of the group-wise measure used with it.

For each case study, I used semantic similarity in four different settings:

Baseline This is a collection of measures instead of a single one. Each measure compares only the concepts from one domain and completely disregards the other domains. This corresponds to the classical single-ontology measure and serves as a baseline to determine whether multi-domain measures outperform single-ontology measures.

Aggregative (raw) This setting corresponds to the aggregative approach, with all the single-ontology values obtained in the baseline setting being averaged with equal weights (eq. 7.1).

Aggregative (weighted) This is the same as last setting, except that the average of the various values is weighted in proportion to the number of annotations in each domain (eq. 7.2).

Integrative This corresponds to the integrative approach. All the ontologies relevant for the similarity calculation are merged into one ontology and then the single-ontology measure is applied to it.

7.2.1 Epidemiology Dataset

Among the annotations for the 228 epidemiology-related papers uploaded to the Epidemic Marketplace, there are annotations made to concepts from the ℕℂ𝕀𝕥 and 𝕄𝕖𝕊ℍ. These two ontologies are much less formal than the rest of the ontologies in this dataset (see Section 2.2 "Ontologies" and **Figure 2.1**). They are also quite broad in their scope. In fact, 𝕄𝕖𝕊ℍ was first introduced as a means to annotate biomedical articles [Rog63] and ℕℂ𝕀𝕥 for annotating cancer-related results [Cor04]; both endeavours need, therefore, a wide range of relevant biomedical concepts.

Table 7.1 – Summary of the annotation in the purged epidemiology dataset. Among the 228 resources in the regular dataset, 24 have annotations only to concepts in MeSH and NCIt and were, therefore, removed from this dataset, resulting in a total of 204 resources. *Cf.* **Table 6.1**.

Domain	Ontology	Coverage	Volume	Diversity
Chemistry	CHEBI	1.2%	1.00	1
Diseases	DOID	57.6%	1.58	45
Environment	ENVO	30.0%	1.00	9
Phenotypes	PATO	1.2%	1.00	1
Symptoms	SYMP	65.6%	3.55	79
Transmission	TRANS	60.6%	1.00	9
Vaccines	VO	29.4%	1.06	16

For example, NCIt covers clinical care, translational and basic research, public information and administrative activities. As such, we can expect that the two vocabularies in fact contribute a bit to all the domains of the epidemiology resources, rather than being specific to one domain. Although these two vocabularies were originally intended to be used in the Epidemic Marketplace only when the other ontologies did not have the necessary concepts, especially as a source of non-biomedical-specific concepts (*e. g.* "Family characteristics", which belongs to the socio-economic sub-domain of epidemiology), they ended up providing the majority of annotations in the dataset (*cf.* the columns "Volume" and "Diversity" in **Table 6.1**).

For these reasons, I calculated semantic similarity and analysed the results in two different ways: first considering all the annotations and second by ignoring the MeSH and NCIt annotations. A third study could have been performed, where the MeSH and NCIt annotations were redistributed among the actual domains they belong too, but this study was not possible, as there is no obvious means to automatically detect which domain each of these concepts belongs to.

While **Table 6.1** contains the statistics for all ontologies, **Table 7.1** contains the statistics for the purged dataset, where MeSH and NCIt annotations were removed, as well as the resources that were only annotated with these ontologies.

Validation of semantic similarity in this dataset was done by determining the degree to which it is possible to predict the diseases (the DOID annotations) from the rest of the annotations. The reason to chose this method was that performing a clinical diagnosis is equivalent to predicting the diseases based on the other known factors (most notably symptoms) and is, therefore, one of the most important problems in biomedical informatics. According to the hierarchy developed in Chapter 4 "Validation strategies" and illustrated in **Figure 4.1**, this validation strategy is classified into the "Classification prediction for single entities" branch.

To this purpose, I used a multi-label machine learning algorithm named ML-KNN, described by Zhang and Zhou [ZZ07] and based on the more general algorithm known as k-nearest neighbours (k-NN). ML-KNN operates approximately as follows:

1. For each resource r, compare it to all other resources using semantic similarity without using the DOID annotations, and find the k resources most similar to r.
2. Build a Bayesian network classifier [FGG97] based on the frequency with which each DOID concept appears in the k neighbours.
3. Use the classifier to calculate the probability that each DOID concept is one of the annotations of r; let $p_r(d_i)$ be the probability associated with concept d_i in resource r.
4. For each resource, sort the DOID concepts according to their associated probability.

Evaluation of this approach can be measured with a number of different methods, making use of the following notation:

- R is the set of all resources in the dataset;
- D is the set of all DOID concepts;
- $C_r \subseteq D$ is the set of DOID concepts annotating r, *i. e.* the set of correct labels; and
- $I_r = D \setminus C_r$ is the set of concepts not used to annotate r, *i. e.* the incorrect labels.

I assessed the performance of each semantic similarity measure using an evaluation measure adapted from Zhang and Zhou [ZZ07] (therein named "coverage"), described as:

$$E = \frac{1}{|R|} \cdot \sum_{r \in R} \frac{|\{d_i \in I_r \mid p_r(d_i) < m_r\}|}{|I_r|} \tag{7.4}$$

where $m_r = \min \{p_r(d_i) \mid d_i \in C_r\}$. This calculates, for each resource r, the fraction of incorrect labels of r that have a low probability associated with it, setting the threshold to the value of the lowest probability of any correct label. In this sense, it is analogous to the specificity at the level of perfect recall: we expect that the number of incorrect labels after the threshold m_r is as high as possible. The perfect similarity measure would completely separate the expected DOID labels from the incorrect ones, resulting in an evaluation $E = 1$.

Other evaluation measures can be applied to this problem. For example, the original ML-KNN paper proposed to measure the fraction of resources for which the most probable label is indeed an expected label, or the fraction of pairs of expected *vs.* incorrect labels where the probability of the correct label is higher than the probability of the incorrect label. All these evaluation measures point to the same conclusions that I present here and, as such, are not shown.

The graph presented in **Figure 7.3** depicts this evaluation measure with respect to the several similarity settings defined in this section, using various values of k in the ML-KNN algorithm. As can be seen, the integrative approach always outperforms the other settings, independently of the value of k, thus showing the superiority of multi-domain semantic similarity in this dataset over single-ontology measures. We can observe that the single-ontology measure performed on the SYMP baseline (using the symptoms ontology) is the most successful baseline.

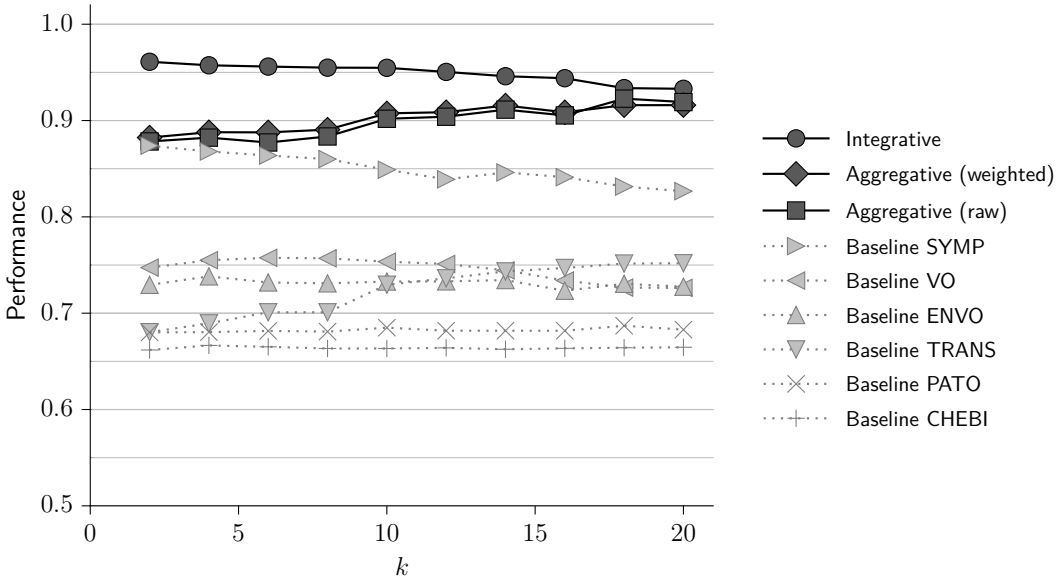


Figure 7.3 – Semantic similarity in the purged Epidemic Marketplace dataset. These results show the performance of the semantic similarity measures using the various settings detailed in the beginning of this section. This graph was obtained using the purged dataset, *i. e.* excluding MeSH and NCIt annotations. Baseline settings are presented as dotted grey lines, and the multi-domain settings as black solid lines.

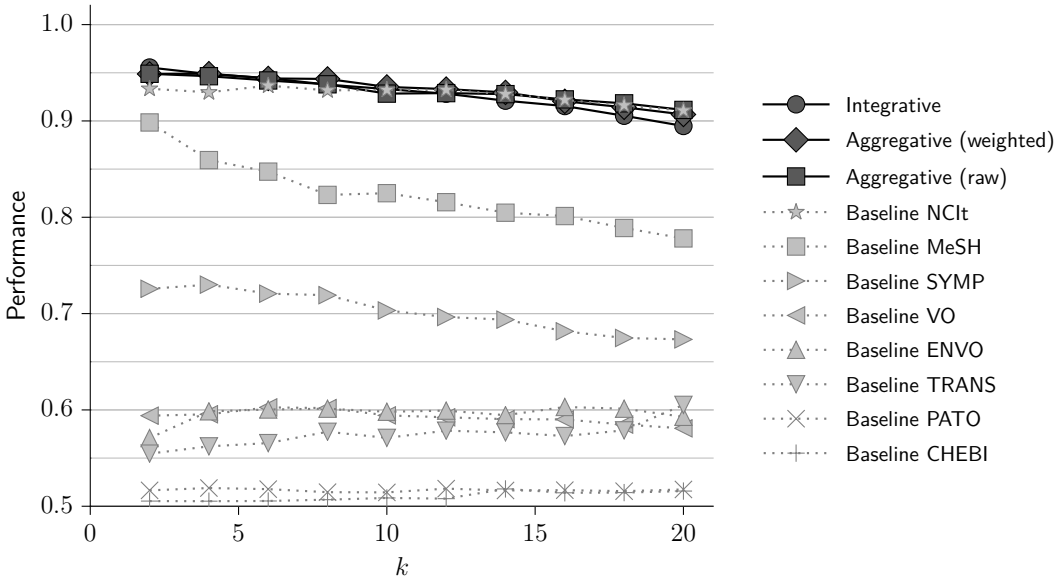


Figure 7.4 – Semantic similarity in the raw Epidemic Marketplace dataset. These results show the performance using the various settings detailed in the beginning of this section. This graph was obtained using all annotations, including MeSH and NCIt. Baseline settings are presented as dotted grey lines, and the multi-domain settings as black solid lines.

This is justified by considering the annotation profile shown in **Table 7.1**. In fact, except for DOID, this is the domain with the highest coverage, volume and diversity of annotation.

Additionally, from the set of domains used to annotate these resources, symptoms are the most closely related to diseases. This baseline shows a performance comparable to the **Aggregative** approaches, especially for low values of k, which means that the other domains have little to add to the information already provided by SYMP (the gap increases with the increase of k). However, the **Integrative** approach shows a higher performance than the other multi-domain approaches for all values of k. This may be justified with the fact that many resources are annotated with different sets of domains. For example, consider the calculation of similarity between a resource that has annotations from TRANS with a resource with TRANS, SYMP and VO annotations. In the **Aggregative** approaches, only the TRANS domain can be used, which results in the method ignoring some annotations; but the **Integrative** approach uses all annotations, irrespective of domain, thus being able to more correctly discern between the resources and being, in general, more accurate. This appears to be an especially relevant result in this dataset, as the coverage of the various domains is small.

Figure 7.4 contains an equivalent graph, obtained in the regular dataset (including MeSH and NCIt annotations). Notice that the baseline performances are not the same as in the previous purged dataset, because this raw dataset contains more resources, specifically more resources that do not have annotations in the domains of those baselines, decreasing their performance. The main conclusion that can be taken from this result is that similarity calculated with multi-domain approaches performs as well as the best baseline (calculated with the NCIt domain). On the one hand, this domain corresponds to the maximum coverage, volume and diversity (see **Table 6.1**), and as such the rest of the domains have little information to add to it. On the other hand, these other domains manage to avoid adding noise to the multi-domain measures: multi-domain measures never show a performance significantly lower than the NCIt baseline. As expected, the best baselines correspond to the two ontologies that span all the domains of annotation.

7.2.2 Metabolic Pathways Dataset

KEGG pathways are manually classified into 33 distinct groups. For example, "Lysine degradation" is classified into the "Amino acid metabolism" group, since Lysine is an amino acid. I exploited this classification in evaluating the performance of semantic similarity in the metabolic pathways dataset by using the similarity values to predict that same manual classification. As above, this is a strategy that belongs to the "Classification prediction for single entities" branch of the validation hierarchy. Specifically, I applied the machine-learning algorithm k-nearest neighbours (k-NN) to predict the class of each pathway given its similarity with the other pathways. The algorithm can be approximately described as follows:

1. For each pathway p, compare it to all other pathways using semantic similarity, and find the k pathways most similar to it. Call this group of neighbour pathways $Q_p = \{q_1, \ldots, q_k\}$.
2. Find the class of each of these neighbours, $C(q_i)$ for $i = 1, \ldots, k$.
3. For each class c' among all the 33 above, count the number of neighbour pathways in Q_p that are part of that class: $f(c') = |\{q_i \in Q_p \mid C(q_i) = c'\}|$.
4. Select $c = \arg\max_{c'} f(c')$, *i. e.* the most common class among the k neighbours. In case of a tie, the selected class is the one with a higher sum of the similarities between p and the pathways in that class.
5. Each pathway is then classified as "correct" if the selected class corresponds to the real one, and "incorrect" otherwise.
6. Performance is reported as the fraction of correct pathways.

As can be seen from **Figure 7.5**, performance was calculated for various values of k, and in almost all cases we observe that multi-domain settings outperform single-ontology ones.

A conclusion that can be taken from this figure is that the different domains perform differently from one another, with the "GO terms" baseline performing on par with the **Aggregative(raw)** multi-domain setting. I refer the reader again to **Table 6.2** (page 87), which summarises the annotations in each domain. For example, only 42.4% of the pathways contain information on drugs, which means that only about 18% of all the pairs of pathways can use this information. Thus, using this domain results in low performance, as expected. On the other hand, all pathways are annotated with GO concepts, with a volume of about 30 concepts per pathway. This, coupled with the fact that semantic similarity in the biomedical domain has been initially explored in GO and has been since thoroughly studied in this ontology more than in the other ontologies, means that it is not surprising that there is little improvement when going from the GO baseline to the multi-domain approaches. Nonetheless, improvements are observed in multi-domain settings for most values of k.

Unlike what happens in the previous dataset, the **Integrative** setting does not clearly outperform the other multi-domain measures. I justify this observation with the fact that the domains in this dataset are more well-balanced than in the previous one, since a majority of the pathways are annotated in a significant portion of the domains (70% of the pathways have annotations in 3 or 4 domains). The "Participants" baseline also shows a high accuracy: it was already established that semantic similarity is a useful technique in CHEBI [FC10; FHC13]. While the accuracy for the "Metabolites" and the "Drug" baselines is low, coupling this information with the other domains increases the performance with respect to the best baseline, namely in the **Aggregative(weighted)** and **Integrative** approaches.

In cases like these, where the domains are balanced in terms of number of annotations, and are known to produce good results with semantic similarity, I anticipate that there is no way, short of actually evaluating the results of semantic similarity, to determine which of the

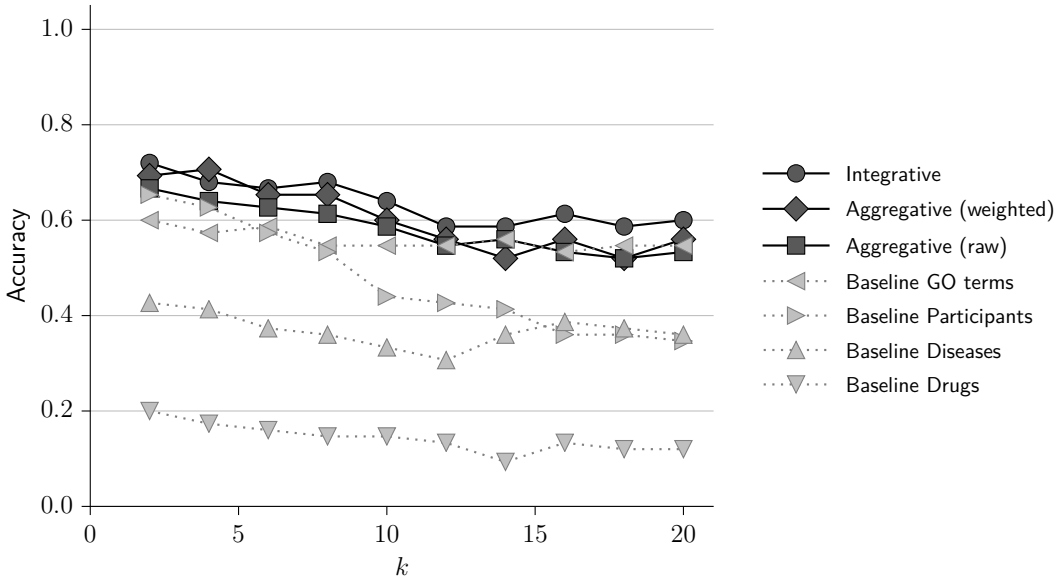


Figure 7.5 – Semantic similarity in the Metabolic Pathways dataset. These results show the fraction of pathways correctly classified, using the various settings detailed in the beginning of this section. Baseline settings are presented as dotted grey lines, and the multi-domain settings as black solid lines.

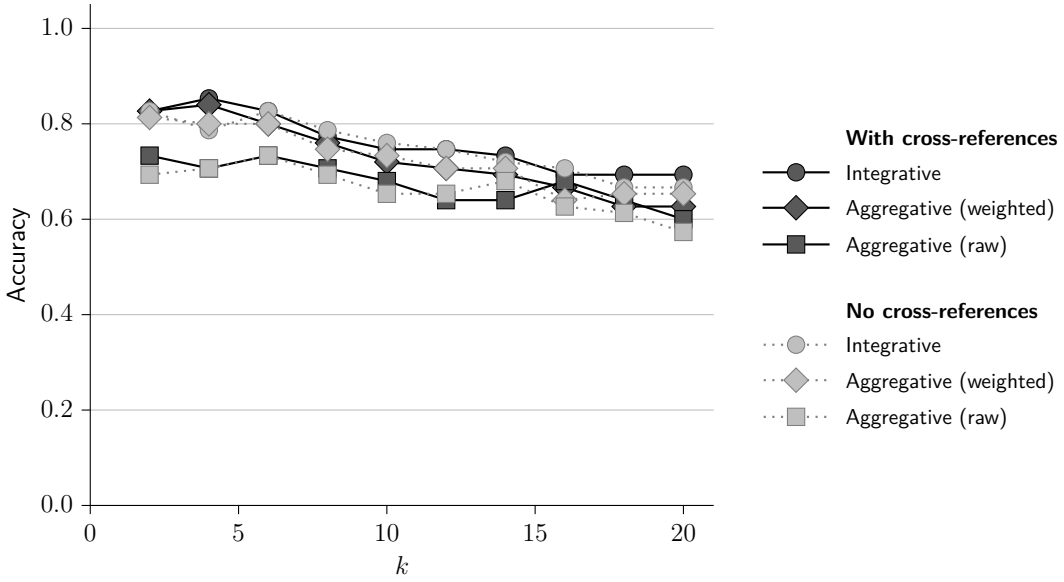


Figure 7.6 – Effect of cross-references on the performance of semantic similarity. As before, these results show the fraction of pathways correctly classified, but now using only the multi-domain approaches. The black solid lines show the results obtained using the cross-references that connect the GO and CHEBI ontologies; the dotted grey lines represent the results obtained without those cross-references. Only the $\text{rel}_{\text{Ferreira}}$ measure can make actual use of the cross-references, and as such this graph shows the results obtained with that measure.

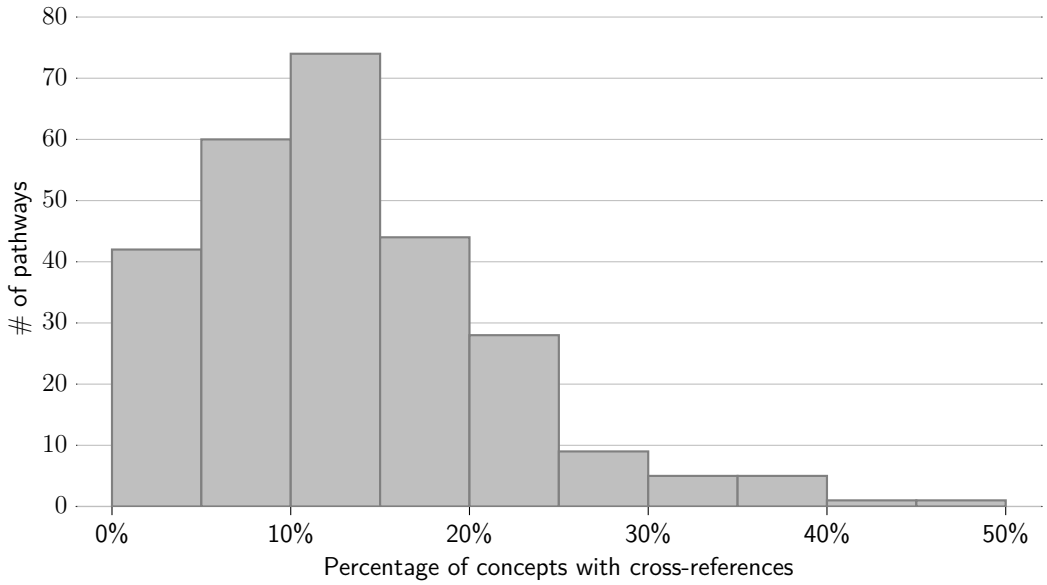


Figure 7.7 – The distribution of the percentage of annotations that have cross-references. This histogram plots the amount of pathways, among the 269 in the Metabolic Pathways dataset, according to how many of their annotations have inter-ontology cross-references.

multi-domain approaches will lead to the best performance. Even so, I expect that at least the integrative approach will always outperform the best single-ontology baseline.

As hinted before, it is also relevant to study the effect that multi-domain cross-references have on semantic similarity. Since $\text{sim}_{\text{Resnik}}$ is agnostic to cross-references, I used instead my own measure of relatedness, $\text{rel}_{\text{Ferreira}}$ (Section 5.2 "Semantic relatedness measure"). Given that using the links between ontologies means that the semantic relatedness algorithm has more information to work with, I originally expected that measures using cross-references would have a higher performance. **Figure 7.6** shows the reverse: adding cross-references to the store of information accessible to the algorithm has no significant impact on the performance. The main reason for this result seems to be that the amount of cross-references is still low, when compared to the size of the dataset and the size of the ontologies. For example, cross-references exist only in $\mathbb{GO}$ ontology, and only for 25% of its concepts (corresponding to only approximately 11% of the concepts of all domains). **Figure 7.7** shows that in the majority of pathways, only 15% of the annotations have cross-references.

On the other hand, this relatedness algorithm has been validated in $\mathbb{FMA}$ measures, which may also contribute to it not being able to use the information from $\mathbb{GO}$ cross-references. In fact, while $\text{rel}_{\text{Ferreira}}$ outperforms $\text{sim}_{\text{Resnik}}$ by a large amount (*cf.* the results in **Figure 7.5** and **Figure 7.6**, where we can see that the **Integrative** approach has accuracy values ranging in the interval $(0.6, 0.7)$ for $\text{sim}_{\text{Resnik}}$ and in the interval $(0.7, 0.8)$ for $\text{rel}_{\text{Ferreira}}$), it seems to be unable to use the cross-references to improve its results even further.

7.2.3 Biochemical Models Dataset

To evaluate semantic similarity in this dataset, I asked a Systems Biologist (Dr Bernard de Bono from University London College) to evaluate a predetermined set of 250 pairs of biomodels, each according to how similar the two biomodels in the pair are (see Section 6.3 "Biochemical Models Dataset" and **Figure 6.2**).

To assess the performance of each semantic similarity measure, I evaluated the degree to which the measures reflect the manual similarity values, an approach that is classified, according to the hierarchy in **Figure 4.1**, as a "Correlation with a manual anchor measure". Since the gold-standard values are not continuous but rather ordinal, this correlation cannot be measured with Pearson's correlation coefficient, but should instead be measured with non-parametric coefficients such as Spearman's rank coefficient or Kendall's τ coefficient. The results shown in this section correspond to Spearman's rank coefficient, but the ones obtained with Kendall's coefficient are equivalent in all aspects.

As can be seen from **Figure 7.8**, the integrative approach outperforms all the other settings, namely the single-ontology ones. In this case, it is even more interesting to compare the results obtained with $\text{sim}_{\text{Resnik}}$ *vs.* $\text{rel}_{\text{Ferreira}}$ (see **Figure 7.9**). The most noticeable difference is that the **Integrative** approach achieves a much higher performance with $\text{rel}_{\text{Ferreira}}$ than with $\text{sim}_{\text{Resnik}}$, even though the single-ontology baselines and the aggregative approaches do not exhibit as large an increase. This seems to indicate that this multi-domain approach is able to thoroughly explore the multidisciplinarity with this measure in a way that it cannot with $\text{sim}_{\text{Resnik}}$.

Like in the previous dataset, GO-based semantic similarity performs highly (even higher than the **Aggregative** approaches). I argue that this happens for the same reasons: semantic similarity has been studied to a higher degree of detail in GO than in the other ontologies, and the coverage and volume of GO annotations is higher than the rest of the domains. In particular, semantic similarity in FMA yields a performance of 0 because, even though 11 models contain FMA annotations (see **Table 6.3**), the gold standard never pairs one of these 11 models with another one of them.

Like in the first dataset, there is a discrepancy between the three multi-domain settings, as the integrative approach outperforms the **Aggregative** ones. I believe the reasons for this are equivalent to the ones presented before: in the **Aggregative** approach, the final similarity score will be an average of four measures, three of which exhibit a low performance. It is not surprising, therefore, that its performance does not increase much with respect to the "GO" baseline (in fact, performance decreases for the **Aggregative (weight)** approach with $\text{sim}_{\text{Resnik}}$).

Again in this dataset, it can be observed that using cross-references does not produce any significant difference. In this case, in fact, I do not present a new figure as it would be so similar to the ones already presented that only a reader willing to use a ruler would be able to tell the difference in the height of the bars. The differences are instead presented in **Table 7.2**.

One other aspect to consider in this case is the overall low correlation coefficient obtained

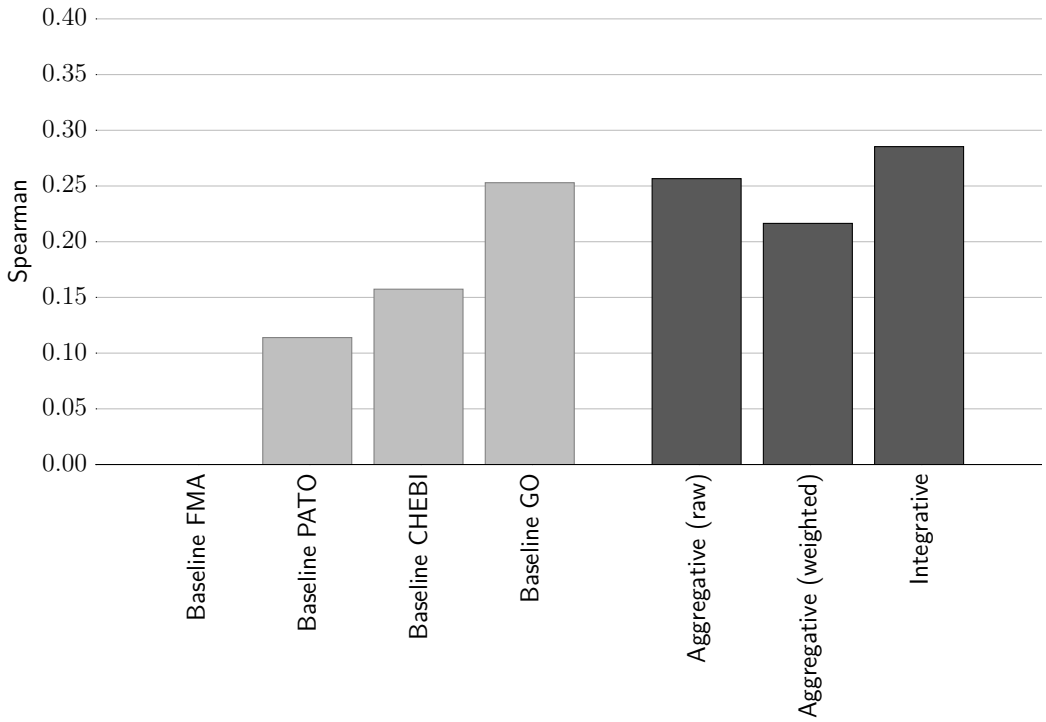


Figure 7.8 – Semantic similarity in the Biochemical Models dataset with $\mathbf{sim_{Resnik}}$. These results show the Spearman's rank coefficient between each semantic similarity measure and the gold-standard, measured with $\text{sim}_{\text{Resnik}}$. In general, the multi-domain measures (shaded in a darker tone of grey) outperform the single-ontology ones (shaded in a lighter tone of grey).

with any of the measures and the multi-domain approaches. The best performance in all cases is obtained with $\text{rel}_{\text{Ferreira}}$ with the **Integrative** approach, but this corresponds only to a correlation of 0.352780. Although not entirely statistically significant, I also calculated Pearson's correlation coefficient for these results, and obtained a value of 0.581401 for this case. A work by Hauke and Kossowski [HK11] shows that when Pearson's correlation coefficient is higher than Spearman's, then there is a (weak, at least) linear correlation between the two variables being measured (as opposed to other types of correlation): in this case, semantic similarity with the $\text{rel}_{\text{Ferreira}}$ measure using the **Integrative** multi-domain approach, and the manual similarity values assigned by the Systems Biology expert. But a linear correlation with Pearson's coefficient of 0.581401 is still a low value. As such, the proposed measures seem to still be lacking in some way, as they do not properly reflect the expert assessment of similarity.

Fortunately, this result does not interfere with my hypothesis. The results presented in the figures of this section can still be used to show that multi-domain measures outperform single-ontology ones.

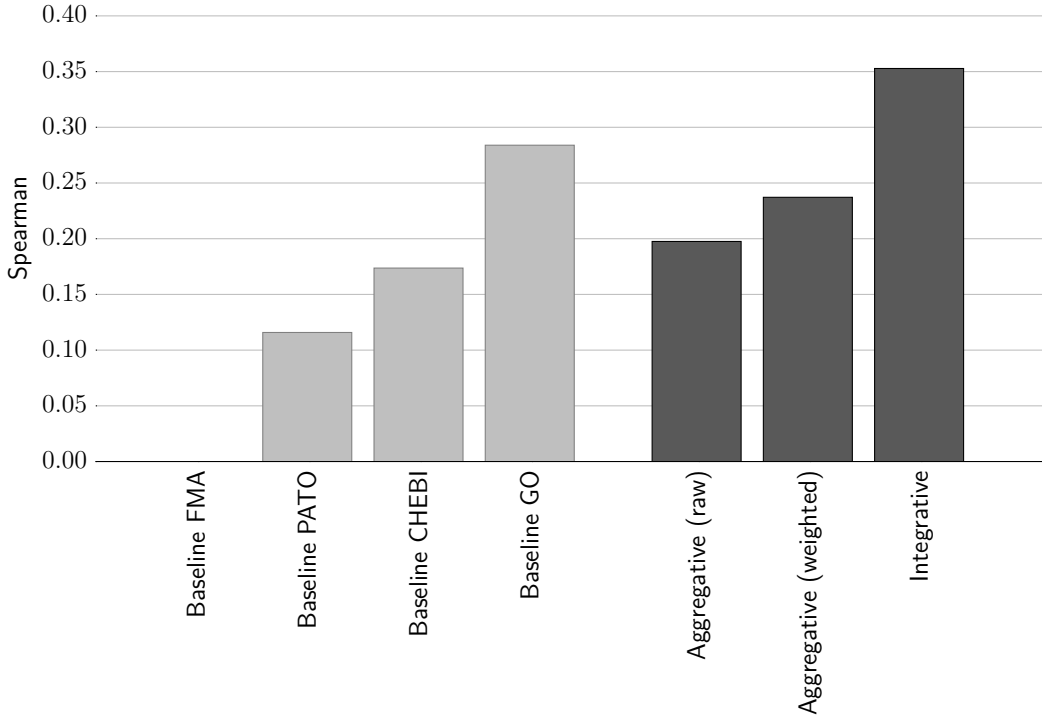


Figure 7.9 – Semantic similarity in the Biochemical Models dataset with $\mathbf{rel_{Ferreira}}$. These results show the Spearman's rank coefficient between each semantic similarity measure and the gold-standard, measured with $\text{rel}_{\text{Ferreira}}$ relatedness measure instead of $\text{sim}_{\text{Resnik}}$. In general, the multi-domain measures (shaded in a darker tone of grey) outperform the single-ontology ones (shaded in a lighter tone of grey).

Table 7.2 – Effect of cross-references on the performance of semantic similarity in the Biochemical Models dataset. These results show the Spearman's rank coefficient between semantic similarity and the gold-standard, measured with $\text{rel}_{\text{Ferreira}}$ relatedness measure. Each column reflects the results obtained without and with cross-references.

Measure		No cross-references	With cross-references
Baseline	FMA	0.000000	0.000000
	PATO	0.115841	0.115841
	CHEBI	0.173646	0.173358
	GO	0.283925	0.286247
Aggregative (raw)		0.197591	0.196394
Aggregative (weighted)		0.237164	0.239875
Integrative		0.352780	0.359541

Chapter 8

Semantic similarity software suite

Working with Web Ontology language (OWL) files is not easy, especially when doing many small things a large number of times. Algorithms that depend on ontology information, such as semantic similarity, cannot be expected to load a set of potentially very large OWL files just to find specific facts represented therein, like what the superclasses of a given class are.

To satisfy the requirements of such algorithms, I propose a solution that enables programmatic access to the information contained in an ontology that does not depend on reading and parsing the ontology for every set of requests, but instead provides *random access* to the elements of the ontology, including not only the concepts but also the axioms stated between them. The main idea of this solution is to insert *useful* information from the OWL files into an SQL database, which can then be queried by semantic similarity algorithms. This software is called `OWLtoSQL`.

Apart from this storage mechanism, I also produced and released a piece of software responsible for computing semantic similarity between concepts and annotated entities in a manner that *i*) depends on the information stored by `OWLtoSQL`, and *ii*) uses a flexible model that can be quickly used by any developer to add their own semantic similarity algorithms. This is called the Multi-Ontology Semantic Similarity (`MOSSy`) tool.

8.1 `OWLtoSQL`

> “ `OWLtoSQL`’s source code is available at https://github.com/jotomicron/OWLtoSQL. ”

Biomedical ontologies are distancing themselves from the simple “hierarchy of concepts” model and are becoming increasingly more complex. As we saw in the previous chapters, ontologies contain disjointness information, existential quantifications, and even other types of axioms. For example, FMA contains the axiom

$$\mathsf{Heart} \sqsubseteq \exists\, \mathit{has\text{-}part}\,.\,\mathsf{Aortic\ valve}$$

which means that for each Heart, there is some Aortic valve to which the heart is related by means of the property *has-part*; in less technical jargon, this means that all hearts have one aortic valve.

Properties themselves are also related to each other, *e. g.* *negatively-regulates*, a property of GO that relates proteins with the processes that they inhibit, is a sub-property of *regulates*.

On the one hand, this increased expressiveness leads to an increase in the richness of the information that can be represented in an ontology, ultimately allowing for a more faithful representation of the reality in a machine-understandable manner. On the other hand, this richness implies a certain complexity in the parsers of the language and an increased difficulty in extracting information from an OWL file in a *random access* way. For example, finding the parts of the Heart, as represented in FMA, is a two-step task:

1. open and parse the fma.owl file into RAM-accessible data structures; and
2. query the data-structures for the necessary information.

Several APIs (Application Programming Interfaces) have been created to perform these steps, such as Jena [Car04] and OWL-API [HB11]. Once opened, the extraction of information is mostly as a small sequence of look-up operations in RAM accessible hashmaps. However, for the one-time random access to the information, this solution is not suitable, as step 1 is time consuming and does not scale with the size of the ontology (to open and parse the aforementioned fma.owl file, a regular-size personal desktop computer — four 2GHz CPUs and 4GB of RAM — takes up to 30 seconds). Having such a large waiting time for one semantic similarity request is highly undesirable: *e. g.* a web service that computes semantic similarity between annotated entities and that takes 30 seconds to return the similarity between Heart and Trachea will likely fail to be adopted by the community.

In order to solve this problem, there are two different possible avenues.

The first approach is to open the OWL file within a computer process capable of inter-process communication, which can then answer client questions like the one about the branches of the Trachea. This approach is quite flexible, since, given an appropriate protocol for the communication between server and clients, it allows the query of any OWL question. However, it is difficult to implement: the communication protocol between the processes needs to align with the OWL specification to ensure that all OWL-valid constructions can be queried and answered. While promising, this idea has yet failed to deliver fully functional software: the only existing implementation I know is OWLlink [Lie08], which is still lacking some useful features. For example, it does not support asking for the names of the concepts, and it is not fully aligned with the latest OWL specification. Additionally, development has been stalled since August 2011.

The second approach is to open the OWL files and extract its information into a more accessible medium, such as a database. OWL can be fully serialised in RDF (see Section 2.4 "Semantic web"), and as such RDF triple stores are an intuitive candidate for storing the OWL ontology, such as OpenRDF Sesame, Jena and OpenLink Virtuoso. These programs are, however, more suited for dealing with contexts where the ontology information is used to reason about the existing data, *e. g.* to infer the type of some instances based on the properties asserted about those instance; instead, my work in semantic similarity is mostly related to querying the ontology itself:

examples of question that are in realm of semantic similarity include "what things are part of a Heart?" and "what are the superclasses of Aortic valve?". Additionally, querying over triple stores is usually done with SPARQL queries [HS13], which is an OWL-agnostic language. While efforts have been made in order to introduce OWL-aware query languages, such as SPARQL-DL [SP07] and SQWRL [OD09], they have not been implemented in any of the existing triple stores.

Furthermore, even a fully working triple store solution equipped with an OWL-aware query language can be slower than necessary for some user needs. For example, asking for all the leaf concepts (concepts which have no subclass of their own) may be a time consuming task, since it involves querying, for each concept, if it has any subclasses. While reasoners can alleviate this task by allowing certain queries to be performed faster (*e. g.* reasoners build a static hierarchy that allows a quick answer to queries like "Is A a subclass of B?"), some queries may still take a long time to run. In fact, a reasoner does not give the number of hypernymy relationships between two concepts, but only whether such a path exists. As such, it can be argued that some algorithms would benefit from a type of caching that computes a single time and stores the information they need in an easy-to-use, fast, and random-access back-end, which is not yet available.

In this context, it is relevant to introduce the idea of *useful* information: presumably, an information-intensive algorithm such as semantic similarity has a static set of information requirements, an *a-priori* established set of axiom types that are needed for the algorithm to work. For example, some semantic similarity algorithms need to know the superclasses of a given class and "how far" the superclass is to the class itself in the hierarchy. As such, and for the purpose of this discussion, I define *useful* information as the total information that an algorithm needs to extract from the ontology in order to run without parsing the original OWL file and without performing time- and resource-consuming computations.

Based on this idea, I developed `OWLtoSQL`, an extensible Java program that is responsible for reading OWL files using the OWL-API and saving any useful information into an underlying relational MySQL database. This allows random-access to any information that is encoded in the original ontology without the need to parse the ontology file again. Indices created on the stored tables guarantee fast retrieval of this information.

This is not the first time that a proposal like this has been made. Zhou et al. [Zho06] and Henß et al. [Hen09] proposed two previous solutions that tried to map all the OWL specification into a back-end database. These solutions, however, are buggy (I have in fact personally approached one of these authors requesting assistance with a bug I experienced, but their response was that they knew about the bug I was seeing, that it was due to a third party library, and that they could not offer further help), outdated (development has been stalled at least since 2010) and do not allow the insertion of non-standard information in the database (such as the edge distance in the class-subclass hierarchy). As such, I had to roll out my own solution.

Other papers have explored the idea of using OWL to build SQL databases, but they use OWL as a means to develop the database *schema*, not to populate a database with the information encoded in the ontology [*e. g.* AKK07; ZK14].

8.1.1 The software model

From the point of view of its user, the interface of `OWLtoSQL` is a configuration file that provides:

- the settings needed to connect to the MySQL database;
- the list of ontologies that are to be loaded in memory and whose information is to be stored in the database; and
- a list of *extractors*, which are Java classes that are responsible for extracting the information from the memory-accessible data-structures containing the ontologies and putting that information in the database.

`OWLtoSQL` then *i*) opens a connection to the MySQL server, *ii*) loads the ontologies using OWL-API, and *iii*) *blindly* executes the code of each specified extractor. Each extractor is an implementation of the abstract Java class `Extractor`, which provides convenience methods to access the MySQL database and the configuration options. In this manner, the definition of *useful* information is provided by the user as the set of extractors to run (see **Figure 8.1**).

Importantly to the idea behind `OWLtoSQL` is the notion that it is highly extensible. Anyone with knowledge of Java and the OWL-API can create their own Java class that extends `Extractor` and then use it as a *plugin* to `OWLtoSQL`.

8.1.2 Configuration file

`OWLtoSQL` reads from a configuration file that the user is responsible for generating, which enables the user to choose the ontologies to load and extractors to run. The configuration file is written using JSON format and it must contain the following elements:

- `"mysql"` determines the host, database name, user name and password needed to connect to the SQL server.
- `"ontologies"` is a list of strings, each containing the `url` that determines where the ontology should be loaded from. These options expect valid URLs that can be used by the OWL-API: as such, URL schemas such as `file:` or `http:` are supported.
- `"extractors"` contains a list of extractor specifications. These are themselves JSON objects containing a `"class"` element that points to the binary name of an extractor Java class and any additional elements that are used to tune the extractor's behaviour. In the case illustrated in **Listing 8.1**, the parameter `"properties"` contains the standard label property defined by the W3C committee [GB14]. Interpreting these options is the responsibility of the extractor class.

More than one extractor can be given (as exemplified in **Listing 8.1**), and the order in which they are given in the configuration file is the order in which they are executed by `OWLtoSQL`.

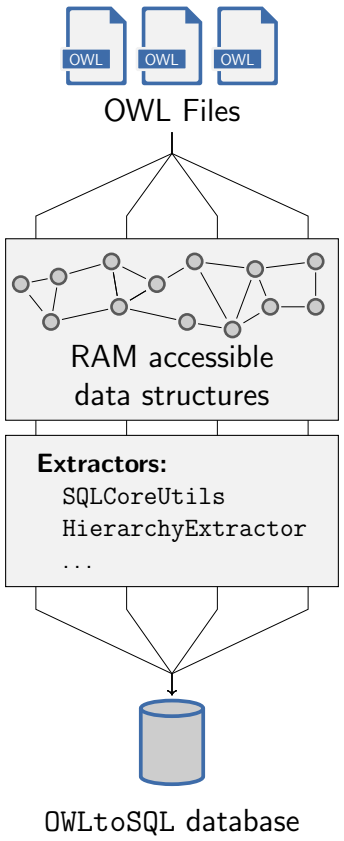


Figure 8.1 – Operation model of `OWLtoSQL`. Several OWL files can be used by this software, which opens and parses them into memory-accessible data-structures (usually hashmaps that allow quick data lookup). Several Java classes extending a common interface named `Extractor` are then used to extract information from these data-structures and store it in the underlying database.

8.1.3 Built-in extractors

Even though `OWLtoSQL` provides the possibility for user-defined extractors, it provides a significant number of built-in extractors. I mention five of them, as an illustration.

The most important one, which is fundamental for the proper functioning of this software, is named `SQLCoreUtils` and it is responsible for extracting the bare minimum information from the OWL files: the entities represented in each loaded ontology. Each entity is stored in a master table and a unique integer identifier is provided to it. This identifier is more suitable for database management than the actual Internationalized Resource Identifier (IRI) (see Section 2.3 "Web Ontology Language") that ontologies provide, as integers can be more easily indexed than the IRI of the entities, which are text that is, in principle, not bounded by a maximum length. The table saves the identifier, the IRI of the entity, and the type of OWL entity that it represents. For example, if an ontology makes use of the concept http://www.w3.org/2002/07/owl#Thing, this entity is stored in the database as an `OWLClass` associated with that IRI, and a unique integer identifier is assigned to it.

```
{
    "ontologies": [
        "http://purl.obolibrary.org/obo/go.owl"
    ],
    "mysql": {
        "hostname": "localhost",
        "database": "db_name",
        "username": "user",
        "password": "passwd"
    },
    "extractors": [
        {
            "class": "pt.owlsql.extractors.NamesExtractor",
            "properties": [
                "http://www.w3.org/2000/01/rdf-schema#label"
            ]
        },
        {
            "class": "pt.owlsql.extractors.HierarchyExtractor",
        },
        {
            "class": "pt.owlsql.extractors.LeavesExtractor",
        }
    ]
}
```

Listing 8.1 – A possible `OWLtoSQL` configuration file. These examples shows the configuration that needs to be provided to `OWLtoSQL` in order to store the information represented in the Gene Ontology, and specifies that the information to extract is the labels of the concepts, the class-subclass hierarchy and the set of leaf concepts.

`HierarchyExtractor` builds and stores the full class-subclass hierarchy of the loaded ontologies. Take for instance the small ontology illustrated in **Figure 8.2**; this extractor stores in the database the facts "Wolf *is-a* Mammal" and "Cow *is-a* Mammal", which are direct class-subclass relationships, but also facts like "Wolf *is-an* Animal", which can only be obtained by traversing two relationships. As such, along with each fact, this extractor stores the minimum distance between the two classes.

`LeavesExtractor` creates a table that contains the leaves of the ontologies, *i. e.* the classes that have no subclass. This extractor depends on the information stored by the previous one, and as such must always be executed after the previous one.

`ICExtractor` calculates the information content of each concept (see Section 3.3 "Node-based approaches", specifically eqs. 3.2 and 3.3). Four different algorithms are calculated and stored in the database.

`NamesExtractor` stores the names of the concepts of the ontologies in the database. By default, this extractor uses the property *rdfs:label* to find the names of entities, but it is possible to specify a different property in the configuration file (see **Listing 8.1**).

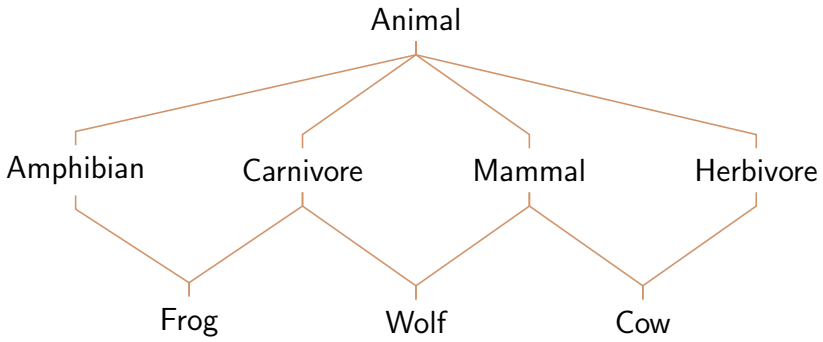


Figure 8.2 – A toy ontology representing some animals. The classes are organised in a class-subclass hierarchy, where the maximum distance between any concept and the root is 2.

8.1.4 Retrieval from the database

The retrieval of information from the database is done directly from the database. This requires that the users of the `OWLtoSQL` back-end are familiar with how the information is stored in the database.

While there are some disadvantages to this (the information is stored in a non-standard format), proper documentation can mitigate these aspects. On the other hand, this allows any programming language that has a MySQL driver to access the information.

8.1.5 Conclusions

`OWLtoSQL` converts OWL files into random-access information in a MySQL database, allowing researchers to streamline work flow that depends on the axioms provided by these files. It is configured with a JSON file that describes the type of information that is needed for downstream applications and provides a series of built-in extractors to actually convert OWL axioms into MySQL tables and rows. Furthermore, it is extensible, which means that Java programmers can create extractor classes that take care of information not already covered by the built-in extractors.

In this sense, I think `OWLtoSQL` contributes to the current panorama in the web semantic by giving ontology users the power to more easily access the axioms they contain.

8.2 MOSSy

> “MOSSy’s source code is available at https://github.com/jotomicron/MOSSy.”

OWLtoSQL stores OWL axioms and other pre-computed information in a database, providing random-access to it. This enables semantic similarity algorithms to exploit the knowledge stored in the database and, therefore, to more quickly perform their calculations. In this section, I present the Multiple-Ontology Semantic Similarity tool (MOSSy), which implements semantic similarity measures precisely by leveraging on an OWLtoSQL database. This dependence means that MOSSy can quickly calculate ontology-based semantic similarity without having the need to wait for the long loading and parsing times associated with reading an OWL file.

Furthermore, by being an open-source tool that can be extended with more similarity algorithms, it can be regarded as a way to reproducibly compare ontology concepts and annotated entities, thus improving the general state of the art in this field of research.

8.2.1 Software model

MOSSy *i*) connects to an OWLtoSQL database, *ii*) reads a configuration file to obtain the set of objects to compare and the semantic similarity algorithm to use, *iii*) performs the comparisons, and *iv*) outputs the resulting similarity values. With this model in mind, MOSSy is actually a thin wrapper around the semantic similarity algorithms. It is coded in Python rather than Java because, given Python’s dynamic nature, algorithms can handle the similarity between two concepts or between two annotated entities. In a typed language like Java, achieving the same goal is cumbersome; in particular, the MOSSy framework would only allow explicitly defined types of entities to be used, while Python allows all possible entity types to be used, due to its duck typing mechanism. On the other hand, Python is one of the languages of choice in bioinformatics, being known and used by 46% of bioinformaticians, contrasting with the 18% that use Java (survey from 2012 [Bar12]).

The real power of MOSSy, similarly to OWLtoSQL, comes from the ability to quickly and easily implement new semantic similarity algorithms that depend on the data extracted by OWLtoSQL. In fact, any Python class that implements a `compare` method that takes two arguments can be specified by the user to perform the comparison. See Section 8.2.4 “Extensibility” for an example. Additionally, the internal MOSSy API contains convenience methods to access the database information, thus accelerating the process of implementing an algorithm even further.

Additionally, given Python’s duck typing mechanism, the user can implement methods to compare not only concepts with other concepts, but any object, as long as the algorithm supports it. For example, we will see in the next section that MOSSy can be used to compare lists of concepts with other lists of concepts, but nothing in the software model needs to be changed to accommodate for that difference, as there is no type requirements for the `compare` method.

Finally, the configuration file is similar to an actual Python script, where the user specifies the comparer and can provide it with parameters if necessary using a familiar syntax.

8.2.2 Configuration parameters

MOSSy expects from the user a configuration file that contains the following information: *i*) the database connection parameters (database, user name and password), *ii*) the semantic similarity algorithm to apply, and *iii*) the pairs of objects to compare. An example of a configuration file is presented in **Listing 8.2**. If OWLtoSQL has been executed to extract the information from GO and CHEBI, this configuration file would produce the output presented in **Listing 8.3**.

```
database = "my_owlsql"
username = "johndoe12"
password = "mypasword"

namespaces = {
    "GO": "http://purl.obolibrary.org/obo/GO_"
    "CHEBI": "http://purl.obolibrary.org/obo/GO_"
}

comparer = simple_model_comparer(
    inner=simgic(ic="seco"),
    aggr=model_avg())

model1 = { "GO": ["GO:0005829", "GO:0008307", "GO:0030049"],
           "CHEBI": ["CHEBI:15377", "CHEBI:16788"] }
model2 = { "GO": ["GO:0035252", "GO:0016266"],
           "CHEBI": ["CHEBI:23588", "CHEBI:15377"] }
model3 = { "GO": ["GO:0035252"],
           "CHEBI": ["CHEBI:23588", "CHEBI:62777"] }

pairs = [(model1, model2), (model1, model3), (model2, model3)]
```

Listing 8.2 – A possible MOSSy configuration file. From top to bottom, this configuration file defines *i*) the database parameters (`database`, `username`, and `password`), *ii*) two namespaces (so that the user can refer to concepts in a more succinct manner), *iii*) the algorithm used to compare the entities (`comparer`), in this case a model comparer that uses sim_{GIC} to compute similarity between lists of concepts and aggregates the values by taking their average, *iv*) the objects being compared (`model1`, `model2` and `model3`), and *v*) the pairs that are to be compared, in this case all the possible pairs are present.

```
model1  model2  0.18541
model1  model3  0.84711
model2  model3  0.21003
```

Listing 8.3 – MOSSy output. This is the output that results from running MOSSy using an underlying database containing the information on GO and CHEBI.

8.2.3 Built-in algorithms

MOSSy already contains several built-in semantic similarity measures, which I divide in three groups:

First, it contains algorithms to compare concepts with concepts. In this category, MOSSy provides the measures proposed by Resnik [Res95] (eq. 3.4), Lin [Lin98] (eq. 3.7) and Jiang and Conrath [JC97] (eq. 3.8). The user can provide parameters to these algorithms (see Section 8.2.2 "Configuration parameters") to specify which hierarchies to use in these measures (*e. g.* GO semantic similarity measures can use a hierarchy containing both *is-a* and *part-of* relationships simultaneously [Lor03]), and whether to include a disjointness factor, according to the description in Section 5.1 "Disjointness axioms in semantic similarity".

Second, it also has algorithms that deal with lists of concepts. On the one hand, it provides the sim_{UI} (eq. 3.11) and sim_{GIC} (eq. 3.12) algorithms, which compare lists of concepts directly; on the other hand, it provides composable approaches that create a similarity matrix with a user-specified concept-to-concept algorithm and then aggregate the several values in the matrix into a single similarity result (see **Figure 3.3**). In this last case, the user can specify exactly which concept-to-concept comparer they want to apply and which aggregation strategy to follow.

Finally, MOSSy provides "model comparison" algorithms, which compare a model with another model. For the purpose of this software, a "model" is dictionary that associates a domain with a list of concepts. This notion is important when dealing with multiple domains: *e. g.* when dealing with biomodels (see Section 6.3 "Biochemical Models Dataset"), it is relevant to use information on enzymes and on chemical compounds, which come from GO and CHEBI respectively. To compare such biomodels with one another it is important to be able to separate GO concepts from CHEBI concepts in different groups. To achieve this, MOSSy expects a dictionary where each key is the name of a domain and the values are lists of terms from that domain (see an example in Section 8.2.2 "Configuration parameters"). In this scenario, a model comparer will use internally a list comparer to compare the GO list of one biomodel to the GO list of the other model (likewise for the CHEBI lists) and aggregates the two values according to some user-specified mechanism, like the average, the maximum or the minimum.

8.2.4 Extensibility

MOSSy facilitates the implementation of new semantic similarity algorithms. The similarity developer needs only to make sure that the information that the algorithm needs to use has been successfully extracted to the underlying OWLtoSQL database. For example, the code in **Listing 8.4** implements an algorithm that returns the *distance* (rather than similarity) between two terms in the class-subclass hierarchy.

Saving this file in a directory that MOSSy recognises is all it takes to enable the user to use the algorithm `edge_distance` in their configuration file.

```
from mossy import sql, utils
from mossy.parse_config import register

@register()
class edge_distance:
    def compare(self, one, two):
        one = utils.get_id(one)
        two = utils.get_id(two)

        sql.cursor.execute("""
            SELECT MIN(h1.distance + h2.distance)
            FROM hierarchy AS h1, hierarchy AS h2
            WHERE h1.subclass = %s AND h2.subclass = %s
              AND h1.superclass = h2.superclass
        """, (one, two))

        return sql.cursor.fetchone()[0]
```

Listing 8.4 – The code of a new MOSSy plugin. This code effectively describes a comparer that calculates the distance between two concepts be counting the number of edges between them in the class-subclass hierarchy extracted to the underlying database.

8.2.5 Conclusion

MOSSy provides a mechanism to allow quick implementation of semantic similarity measures based on OWL ontologies and on the OWL information that has been extracted to an underlying OWLtoSQL database (it does not require loading and parsing OWL files because OWLtoSQL takes care of producing a database where all the needed information is stored). Furthermore, it enables users to easily create a configuration file containing the pairs of objects that they want to compare and the algorithm that they want to use, producing a TSV file with the results of performing the comparison.

Finally, MOSSy contributes to the current panorama in semantic similarity measures, providing a reproducible mechanism to deal with multiple ontologies. Its software model allows quick implementation of measures to be tested, but also allows its use in a production environment.

PART III

Final remarks

> I think and think for months and years. Ninety-nine times, the conclusion is false. The hundredth time I am right.
>
> — Albert Einstein

Chapter 9

Conclusions

Semantic similarity in the biomedical domain has been used to compare entities like proteins, chemical compounds and metabolic pathways. However, biomedical knowledge is intrinsically multidisciplinary: for example, metabolic pathways are related with chemical compounds, proteins, and even other types of concepts such as diseases; epidemiological resources are related to concepts from a wide range of domains, like diseases, symptoms, environmental conditions, *etc.* Comparing these multidisciplinary entities based on their semantics is a problem that, until now, had not yet been tackled.

9.1 Summary of contributions

The main objective of my PhD work was to research methods to handle that problem, *i. e.* I studied and offered solutions to the issue of multi-domain semantic similarity measures. The main hypothesis presented in this document is that multi-domain semantic similarity measures can be constructed by lifting existing single-ontology algorithms into the multidisciplinary case. The main result of this work was, indeed, the empirical proof of this hypothesis by validating the use of multi-domain measures in several biomedical datasets (resources in a epidemiology marketplace, metabolic pathways, and mathematical models of biochemical systems).

While the quantitative results are not best understood in numeric format (in fact most of the results in Chapter 7 "Multi-domain semantic measures" are presented as graphs), it is visible in all those graphs that multi-domain semantic similarity almost always outperforms the single-ontology baselines, with very few exceptions: *e. g.* a baseline using a wide-coverage ontology, such as $\mathbb{NCIt}$, has the same performance as the multi-domain approaches, and the $\mathbb{GO}$ baseline also has a high performance given that the amount of annotations from this ontology far surpasses the amount of annotations in the other ontologies. In all cases, however, the **Integrative** approach to multi-domain semantic similarity always outperforms the best single-ontology baselines. This means that, in fact, the knowledge encoded in the annotations from one domain complements the knowledge encoded in the annotations from the other domains, thus leading to the idea that, in fact, technology that deals with several domains simultaneously needs to be developed in order

to properly explore all the information contained in multidisciplinary resources.

As in all scientific endeavours, getting to prove an interesting and useful statement requires that a significant amount of work be performed "under the hood". In fact, when I started my work, few to no publications existed that dealt with the issue of multidisciplinarity in the biomedical domain and, as far as I was able to ascertain, in any other scientific field. For this reason, the methodology I set for myself achieved intermediate results that were also essential to support the hypothesis.

On the one hand, I created a hierarchy of semantic similarity validation strategies (Chapter 4 "Validation strategies"). This hierarchy, created with a reproducible method, can be used to categorise semantic similarity measures according to the method used by their creators to perform validation, which not only helps categorise the measures, but also allows researchers interested in using semantic similarity to choose a measure that has been shown to have a high performance in the type of problems at hand.

On the other hand, I also contributed to the current panorama in single-ontology semantic similarity (Chapter 5 "Towards OWL-aware similarity"). While this was not a requisite for the ultimate proof of the hypothesis, it provided a first step towards including formal logic constructions in the semantic measures existing today: since my approach to multi-domain similarity is based on pre-existing single-ontology measures, any improvement made to the existing measures will also have a positive impact on the overall performance of the multi-domain measures. In fact, one of the measures developed in this context, $\text{rel}_{\text{Ferreira}}$, is able to calculate relatedness based on on all the properties associated with the concepts of the ontology, not just its class-subclass hierarchy. The success of this measure in the single-ontology world (where it was used to determine whether pairs of anatomical concepts are implicated in the same disease) is reflected in the multidisciplinary datasets, since using it as the base of the multi-domain measure increases the performance with respect to purely similarity measures (such as $\text{sim}_{\text{Resnik}}$).

Furthermore, to properly study multi-domain measures, I had to collect multidisciplinary data, which resulted in a collection of three datasets (Chapter 6 "Multi-domain data"). The first dataset comes from the epidemiology field and contains references to epidemiological articles annotated with concepts from a network of epidemiology-related ontologies. The second dataset contains metabolic pathways annotated with the metabolites that are converted in the pathway and the proteins responsible for catalysing those reactions, as well as the diseases associated with the malfunctioning of the pathways and the drugs that affect them. The third dataset contains mathematical models of biological systems, again annotated with the intervening metabolites and proteins, but also with the anatomical places where those systems are located and the physical quantities measured in those mathematical models.

Finally, given the large amount of information that the measures deal with, I had to develop software mechanisms to cope with the size of biomedical ontologies, namely `OWLtoSQL` and `MOSSy`, which provide an automatic way to run semantic similarity calculations both faster and in a more reproducible manner. As such, this part of my work was more technical than exploratory.

I also participated in several other activities in parallel to the ones described here, which, although not directly related to semantic similarity, were essential to my understanding of how this technique can assist the advancement of science. Appendix B "Auxiliary projects" mentions my work in the Epidemic Marketplace (especially in developing the Network of Epidemiology-Related Ontologies), text-mining, and ontology matching.

As explained in Chapter 1 "Introduction", the results obtained in this work were entirely directed at the biomedical domain, and as such only this area of research directly benefits from these results. However, none of the methods I developed is specific for this area of research and can be adapted to work in other domains. For example, the amount of information living in the semantic web (see Section 2.4 "Semantic web") is also increasing and we will soon need the power of computational methods to deal with that amount of data. From a technical point of view, adapting the presented methods to this area is trivial, as the only requirement is that the necessary ontologies exist and are included in an `OWLtoSQL` database.

9.2 Some shortcomings

Not unlike other scientific endeavours, the work that I performed during my PhD has its own shortcomings and weaknesses. Here I present a few, together with possible avenues to solve them.

I did not find evidence to suggest that using cross-references increases the performance of semantic similarity. I originally hypothesised that semantic similarity measures that use the cross-references would outperform the same measure but without using such links. The results in Section 7.2 "Results" indicate that this is not the case. This can be due to the small number of such references, which are bound to increase both in quantity and in quality as the inter-domain knowledge representation efforts increase. While the long-term solution involves waiting for the proper knowledge to be encoded in machine-readable formats, a short-term solution to this problem is to use external sources of information to find inter-domain links between the concepts in the ontologies of relevance. For example, I propose using text-mining techniques to find co-occurrences of anatomical terms and disease names in scientific literature. Pairs of concepts often mentioned together can then be inferred to have some sort of relationship. Furthermore, we can explore the frequency with which such co-occurrences appear in a corpus to assign a strength for these relationships.

Another weakness of this work is that the results obtained to validate the the multi-domain semantic similarity approaches are not representative of real-world scenarios. For instance, using semantic similarity to classify metabolic pathways in groups that have already been assigned manually shows that the measures are sound, but does not actually produce new knowledge. Additionally, it would be useful to harness the power of online data repository users to establish whether new annotations predicted from existing ones are correct.

A third problem with the results is that, in some cases, performance indicators are not as high as was desired. For example, in the biomodels dataset, the best performance was achieved using

$\text{rel}_{\text{Ferreira}}$ as the group-wise single-ontology semantic similarity with the integrative approach, but these results show a Spearman's rank correlation coefficient of about 0.35. To solve this issue, I expect that tuning the measures will account to an increase in the performance indicators. For example, $\text{rel}_{\text{Ferreira}}$ is a measure that can be tuned with respect to the weights assigned to each property. Running experiments where the weights are changed according to some criteria may be useful to increase the performance. Note, however, that the small absolute correlation does not invalidate the conclusion that multi-domain measures outperform single-ontology ones: this hypothesis still holds, and in fact by a large margin (see *e. g.* **Figure 7.9**).

It is pertinent to notice that these issues can all be solved in time. In fact, as future work, I propose that these are exactly the next steps to develop a fully cohesive multi-domain semantic similarity theory.

9.3 Future work

Besides the points raised in the previous section, there are at least three more aspects that I feel would greatly improve the overall panorama in multi-domain semantic similarity.

I would like to explore the idea of calculating the "relevance" of concepts to assist the computation of $\text{rel}_{\text{Ferreira}}$. This relatedness measure first finds the semantic neighbourhood of a concept c by traversing the properties between concepts and building a graph centred around c, and doing it recursively. This semantic neighbourhood is, therefore, exponential on the number of "layers" that we want to capture (although in practice the number of layers is relatively small compared to the size of the ontology). To mitigate the effort of this step, we could devise an algorithm that decides whether a concept is relevant, reducing the size of the neighbourhood and therefore the execution time. This relevance measure could also benefit other measures, such as sim_{GIC}, which must know for each concept the set of its superclasses; by storing only the relevant superclasses, we could improve the speed, memory requirements, and accuracy of this measure. I have already obtained some preliminary results on this idea which suggest in fact that accuracy increases when only a fraction of superclasses is considered. Further studies need to be developed, however.

I would also like to understand the effect of the aggregation mechanism that is used to compare lists of concepts with a concept-wise semantic similarity (see Section 3.5 "Comparing annotated entities"). For example, to use $\text{sim}_{\text{Resnik}}$ as a group-wise semantic similarity, I used the Best Match Average (BMA) approach to convert the matrix of similarity values into a single value. BMA does this by finding in each row and column of this matrix the highest value and then averaging the values (see **Figure 3.3**), but a possible alternative would be to use, from each row and column, more than one value using, for example, the T-conorm idea proposed by Lehmann and Turhan [LT12] and already explored in Section 3.5 "Comparing annotated entities".

As mentioned in Chapter 4 "Validation strategies", I also intend to create a tool that assists semantic similarity developers in setting up a validation step to their own measures, based on the

hierarchy described in that chapter; furthermore, the hierarchy itself can be included within one of the already existing ontologies, both as a means to standardise it and as a way to motivate its use by the community and its future extension to accommodate other domains.

Finally, as a matter of speed, I would like to explore and modify the current implementation of `MOSSy` so that it can use asynchronous programming and so that it can reduce its dependency on the underlying MySQL database, for example by storing frequently requested information in a local cache. I expect that such modifications would greatly increase the speed of execution, particularly on a multi-core machine.

9.4 Last thoughts

It is undeniable that science can no longer be performed by human mind alone. The data being produced today is so extensive in size that it has become impossible to be aware of all of it without the assistance of computerised systems that crunch the information and hand it over to the scientists in more manageable formats. Semantic similarity is but one aspect of this whole automatic pipeline, a cog in the machine that intends to assist scientific progress.

My contribution, as most contributions in today's scientific community, is but a tiny bump on the frontier of human knowledge, but it so happens that, along with the millions of other scientists working towards knowledge discovery, it is building and improving our own understanding of ourselves and our world. In this sense, I believe my work is a small but steady step towards the future of science.

PART IV

Back Matter

> The moment a man sets his thoughts down on paper, he is in a sense writing for publication.
>
> — Raymond Chandler

Appendix A

List of ontologies

The biomedical informatics community is highly committed to the machine-readable representation of biomedical knowledge. This is illustrated by the increasing number of ontologies being developed focussed on sub-domains of this vast area of research, as well as their increasing size and quality. In particular, the community has established an ambitious goal to represent all of the relevant knowledge for this domain in ontologies, with projects stemming from this goal such as BioPortal, the OBO Foundry, and OntoBee.

BioPortal is an online platform that provides access to biomedical ontologies [Noy09; Whe11]. As of October 2015, it contains 467 ontologies and a total of almost 6.4 million concepts. These ontologies are related to each other through "mappings", which are community provided alignments between the ontologies: for instance, the concept of Femur from the NCIt is mapped with the relation *skos:closeMatch* to 34 concepts from 26 other ontologies. These mappings express the notion that all of these concepts represent the same real-life idea, *i. e.* they are different (sometimes complementary, sometimes distinct) representation of the upper leg bone.

The OBO Foundry is a collaborative experiment designed with a purpose [Smi07]:

> "To establish a set of principles for ontology development with the goal of creating a suite of orthogonal, interoperable, reference ontologies in the biomedical domain."

As of October 2015, this foundry has created a set of principles to guide biomedical ontology development, and they list 9 ontologies that most faithfully obey them (in domains such as anatomy and molecular function), along with 126 other ontologies distributed through 28 distinct domains of knowledge that try to follow the guidelines but have yet to be accepted as full OBO ontologies.

These two projects have different views on the work needed to release an ontology to the community. While BioPortal is a free store of ontologies, where any user can upload an ontology without approval by any entity, the OBO Foundry is run under the expectation that ontologies must be evaluated by the community before being endorsed and accepted as reference ontologies. Together with the use of the objective guidelines to direct the ontology development process, this ensures a minimal amount of quality that is not guaranteed to be present in BioPortal's

ontologies. In fact, BioPortal's objective is not to be a hub of good quality ontologies but simply as a front-end for users to access them.

Similar to BioPortal, OntoBee works as a front-end to serve ontology requests to users [Xia11]. Its backed by the OBO Foundry ontologies and, as such, it can only answer queries about the concepts of those ontologies.

Outstanding examples of biomedical ontologies that have been regarded by the community as reference ontologies to represent sub-domains of knowledge, and which have been used throughout my work, include:

Gene Ontology (GO) The principal focus of GO is on proteins and other gene products (molecules that are created based on DNA): this ontology contains three branches, one for the biochemical functions of gene products, one for their cellular localization, and one for the biological processes in which they participate. The ontology contains, as of October 2015, over 40,000 concepts, related to one another by means of 8 properties. It has been in development since 2000, the year that the first human genome was sequenced [Ash00].

Chemical Entities of Biological Interest (CHEBI) The focus of this ontology is on small molecules that have a biological role, especially (but not exclusively) in the human organism. This ontology represents over 44,000 concepts, related by means of 9 properties. It has been in development since 2007 [Deg08], and contains information integrated from more than 20 different external sources.

Foundational Model of Anatomy (FMA) This ontology represents the domain of human anatomy. The development of this ontology started in 1995 [Ros95] and has since then gone through several major overhauls. It currently contains almost 80,000 concepts, which are related to one another by approximately 60 different properties. While this ontology has been initially developed using techniques different from the OWL language, it has now been converted to OWL. However, some of the information in the original format is not expressible in OWL and is missing from this version [GZB06; GGD13]. For example, only 6 properties have been ported to OWL.

Human Disease Ontology (DOID) This ontology describes human diseases, in a clinically relevant manner, and includes genetic, environmental and infectious diseases. DOID encapsulates a comprehensive theory of disease. Its structure and external references to other terminologies enable the integration of disparate datasets [Osb09].

These are the some of the ontologies that have been developed with greatest attention to detail in the biomedical domain. They satisfy three characteristics that largely increase their usefulness: *i*) comprehension — most of the relevant concepts for each domain are represented in some way in the ontologies; *ii*) precision — concepts are specifically defined, *e. g.* GO contains

the concept Production of molecular mediator of immune response; and *iii*) detail — the level of detail and granularity in the ontologies is high, *e. g.* ℂℍ𝔼𝔹𝕀 contains the concept Carbon-12 atom, and even subatomic particles, and 𝔽𝕄𝔸 contains the concept Cell.

Apart from these content-wise characteristics, there are other properties that make these some of the most successful biomedical ontologies. First, they are formal, follow first-order logic constructions and are generally deployed in OWL or an equivalently formal ontology language (like OBO). Second, they are community driven, which means they are free to use and publicly available, and, more importantly, provide a minimal guarantee of maintenance. Third, the ontologies are being used by the community to annotate complex entities (like proteins, metabolic pathways, *etc.*).

The formal ontologies used in my work that are not part of the previous list are:

- **Environment Ontology** (𝔼ℕ𝕍𝕆) represents environments and environmental conditions.
- **Phenotypic Quality Ontology** (ℙ𝔸𝕋𝕆) represents qualities that are inherent to concepts from other ontologies, such as gene products or anatomical entities. Examples of qualities are Red, High temperature and Small.
- **Symptoms Ontology** (𝕊𝕐𝕄ℙ) represents human symptoms, which are defined within this ontology as "perceived changes in function, sensation or appearance reported by a patient and indicative of a disease".
- **Transmission Modes Ontology** (𝕋ℝ𝔸ℕ𝕊) represent modes of infectious disease transmission.
- **Vaccines Ontology** (𝕍𝕆) represents vaccine-related concepts.

At last, other vocabularies used in my work are 𝕄𝕖𝕊ℍ and ℕℂ𝕀𝕥. These are not formal ontologies in the sense described in Section 2.2 "Ontologies" — they are hierarchies of concepts that are related to one another with underspecified properties. They *do* have OWL representations that try to capture their hierarchy, but since the same OWL property is used to represent all the relationships between concepts, which are not always equal in semantics, the concepts represented in these ontologies do not accurately reflect reality in a logical manner.

Appendix B

Auxiliary projects

This appendix describes three research efforts where I participated that are tangentially related to my work in semantic similarity. Although none of them contributed to the direct research and development in semantic similarity, they were useful in two senses: *i*) they helped me be more familiar with the practices of knowledge representation, data federation and information sharing; and *ii*) they provided a means for me to be acquainted with particular examples of contexts where application of semantic similarity is used as part of other, bigger systems.

B.1 Semantic web in the Epidemic Marketplace

During the course of one year, I participated in the Epiwork project, an European project that ran from 2009 to 2013, funded by the Seventh Framework Program (FP7). This project aimed at developing the appropriate framework of tools and knowledge to design epidemic forecast infrastructures. The tasks assigned to the LaSIGE partner were:

- to develop the Epidemic Marketplace (EM), a repository of epidemiological data;
- to create a website that serves as the front-end to the repository; and
- to define ways to annotate the repository data.

I participated in this project as an expert on semantic web. Namely, I was in charge of *i*) making the data more accessible from the semantic web point of view, particularly to the other partners and their automatic tools, as well as *ii*) increasing the digital preservation of the resources in the EM. My participation has culminated in two contributions:

- a semantic metadata model designed with the specific needs of epidemiology data in mind, which was used to guide the annotation process of the epidemiology resources [Cou12]; and
- a Network of Epidemiology-Related Ontologies (NERO) representing most of the domains of epidemiology (chemistry concepts, diseases, symptoms, environmental conditions, methods of transmission, vaccines, sociology, geography, *etc.*), to be used as source of concepts in the metadata of the resources [Fer12].

The metadata model defines a set of *slots* that provide data owners specific topics relevant for epidemiological data, which can be used to guide the annotation process. It is based on the Dublin Core, a vocabulary of terms used to describe web resources (*e. g.* video, images, web pages), as well as physical resources (*e. g.* books, music records, artwork) [DCM12]. Being based on a popular standard for annotation is an advantage in three fronts:

1. Most of the necessary information needed to describe a resource already exists (terms such as author, publication date, references, *etc.*); we had to add only epidemiology-specific terms.
2. It increases interoperability.
3. It ensures long term usability, which contributes to the preservation of the data. In effect, while the EM website has been discontinued, due to the lack of funds, the data still exists, as well as their annotations.

The metadata model is divided in three sections [Fer13]: *i*) a technical section that contains terms related to the digital nature of the resources (their unique identifier, the name of the EM user that uploaded the data, the date of submission, *etc.*); *ii*) a general section, containing the non-epidemiology-related information of the resource (title, author, description, creation date, *etc.*); and *iii*) a content-specific section, with terms specific to epidemiology, including information on diseases, symptoms, social conditions, *etc.*

Most of the content-specific metadata is meant to be provided by the user as ontology concept identifiers. Using ontologies to fill the metadata of a resource contributes to its machine-readability, but also to the preservation of the data. Ontologies provide:

– an objective and traceable meaning to the metadata;
– a controlled vocabulary, thus contributing to the interoperability of the metadata with other semantic web systems;
– a language agnostic vocabulary, which avoids the pitfalls of natural language processing; and finally
– support for reasoning and other semantic tools (such as semantic similarity).

To maximise the benefits of using ontologies, I contributed to the establishment of the Network of Epidemiology-Related Ontologies (NERO). This is a multi-domain collection of ontologies that represent parts of the epidemiology field: *e. g.* diseases, symptoms, chemical compounds, modes of transmission, *etc.* It also contain non-biomedical ontologies to represent demography, environmental conditions, geographical regions and socio-economic conditions. NERO was integrated in the Epidemic Marketplace as a way to facilitate annotation of resources, by suggesting concepts based on the content of the resource and facilitating the annotation process with an auto-complete-like feature that suggested concepts from those ontologies. As such, NERO bridges the gap between automatic systems and epidemiology, a domain which traditionally

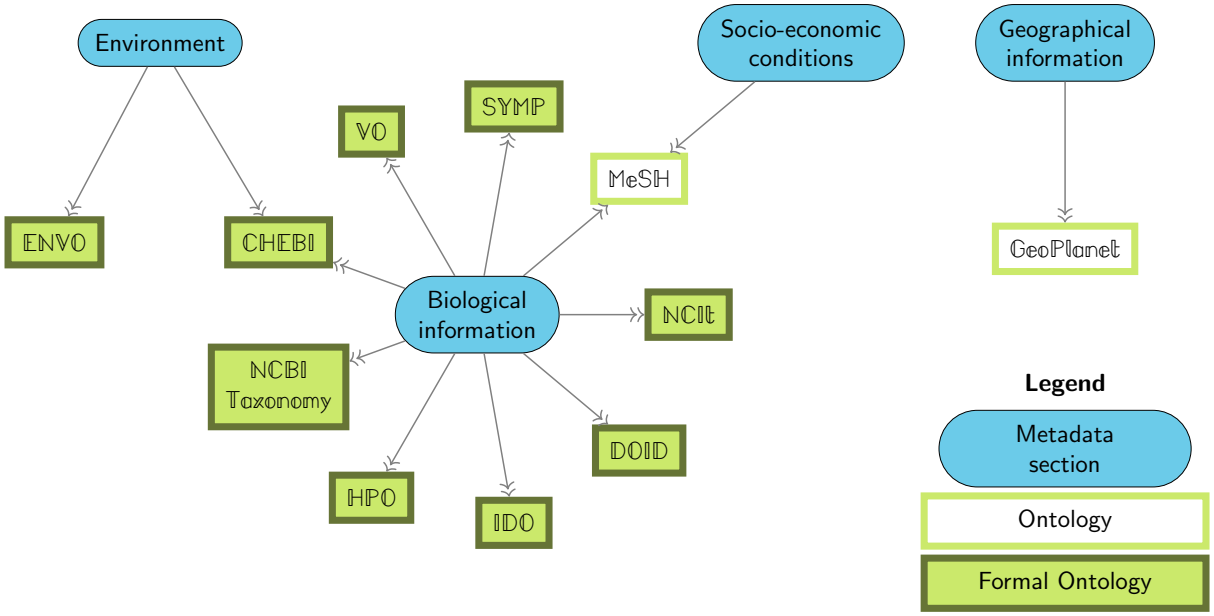


Figure B.1 – The Network of Epidemiology-Related Ontologies. This figure represents NERO as a set of metadata sections (represented as blue round rectangles) and the ontologies used to annotate the epidemiology resources (in straight rectangles). Each metadata section is associated with a set of ontologies that contain the concepts relevant for that section. There is a difference between formal ontologies (represented with the shaded green background) and the other vocabularies, as explained in the last paragraph of Appendix A "List of ontologies".

makes poor use of computer power (possibly because of its high heterogeneity), by bringing the semantic web into it. In fact, prior to NERO, there was not an expressive way to annotate resources with ontology concepts in this field. **Figure B.1** contains a graphical representation of which ontologies are used to fill the metadata model for an epidemiology resource.

B.2 Text-mining

Text-mining aims at extracting relevant information from unstructured natural text. The meaning of "relevant" depends on the actual goals of the text-mining process: for example, automatic news processing systems can perform "Sentiment analysis", a technique that detects whether the opinions expressed in text (in a full article, a blog post, a tweet, *etc.*) is positive or negative; advanced algorithms can even classify text based on more specific emotional states, such as "angry", "sad" or "happy". In the biomedical domain, text mining is an important part of scientific discovery: it can be used to find drug targets and biomarkers, for drug repositioning, to create a clinical overview of a certain therapeutic area, to create domain specific databases, *etc.* [FA15].

The first preliminary study I was part of, in the context of text mining, was the application of semantic similarity to disambiguate geographical names in news articles [Bat12]. Names of geographical features (called "toponyms") are particularly ambiguous: a particular case is in the

name *Lisboa*, which represents up to 41 different locations in the territory of Portugal alone, from streets to a municipality, a city and a region. Being able to properly identify which place is being referred to in text is important to further process that text. One way to achieve this is by:

1. associating each toponym with a set of its possible locations;
2. comparing all the locations within each possible arrangement using semantic similarity;
3. finding the arrangement with highest overall similarity score and choose it as the disambiguated set of locations.

In this work, we used Geo-Net-PT, a geographical ontology of the Portuguese territory, which contains more than 400,000 geographical locations, organised in a hierarchy (*e. g.* Portugal *contains* Lisbon city). This hierarchy can be used to compute semantic similarity with the algorithms mentioned in the main document, thus allowing the disambiguation process above.

I have also contributed to text-mining approaches in the biomedical domain. One of the most important tasks in text processing in biomedical informatics is the identification of entities such as chemical compounds in text. This allows further processing (for example, the detection of interaction between compounds). I have worked as a semantic similarity expert with a set of colleagues in text-mining in this context. In a first step, we investigated whether semantic similarity can be used to disambiguate chemical names in text, just like I had done previously in the geographical domain [LFC15]. In a second step, we investigated whether semantic similarity can also be used to improve the overall performance a system designed to find interactions between two chemical compounds in text [LFC14]. For example, the sentence "*Trilostane* may interact with *aminoglutethimide*, causing too great a decrease in adrenal function" describes the interaction between two compounds (in slanted text), which the system is able to find. Semantic similarity was used here in an effort to reduce false positive interactions found by the system.

B.3 Ontology alignment

Part of my research in multiple-ontology semantic similarity was dedicated to the study of ontology alignment techniques. As explained throughout this document, multi-ontology semantic similarity is enhanced if the ontologies used to compute similarity are related to each other in some way. In general ontology alignment is done by asserting that two concepts from different ontologies are equivalent. In the context of multiple-ontology single-domain measures (see Section 3.6 "Multiple-ontology semantic similarity"), complementary ontologies improve the accuracy of similarity only if the ontologies are *linked*.

For example, with the aligned pair of ontologies in **Figure B.2**, it would be possible to assign a relatively high value to the similarity between Antibiotic activity and Antibacterial role, but only if their parent concepts (Antimicrobial activity and Antimicrobial role respectively) are explicitly marked as *equivalent*, as the dotted line suggests.

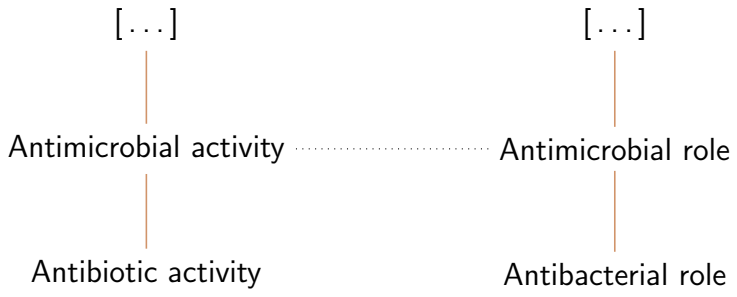


Figure B.2 – Partial alignment between two ontologies of the biochemical domain. The two ontologies partially illustrated here represent the same domain of reality, namely the roles of biochemical molecules. The dotted line represents an *equivalence* link between the two, and can be explored by multiple-ontology semantic similarity to compare concepts in different ontologies.

The set of equivalences between multiple ontologies (called an *alignment*) is, therefore, essential to single-domain multiple-ontology similarity measures. However, finding them is labour-intensive, given the amount of concepts in biomedical ontologies, which has lead the community to develop "ontology matching" algorithms, which find, automatically or semi-automatically, equivalent concepts within two or more ontologies [ES07]. Ontology alignments can be made with simple textual matching, which rely on dictionaries and thesauri to increase their recall (for instance, the fact that "activity" and "role" are synonyms might be used to match the two concepts in the ontologies from **Figure B.2**); they can also leverage on the structure of the ontologies to find related concepts (concepts with many linked subclasses should themselves be linked); and they can also explore the logical definitions on the ontologies to find these matches [SE05].

During the early stages of my PhD, I worked under the assumption that ontology alignment would be essential for my research. As such, I participated in Semantic Ontology Matching using External Resources (SOMER), a project funded by the Fundação para a Ciência e Tecnologia (the Portuguese Foundation for Science and Technology), and which ran from 2012 to 2014. One of the tasks that I developed was the alignment of geographical ontologies (namely the Yahoo! GeoPlanet, a geography ontology for world-wide locations, and the Geo-Net-PT, mentioned above).

As it turns out, ontologies that represent different domains of the biomedical information are usually already quite orthogonal, given the OBO Foundry's principles and the best practices in ontology development (see Appendix A "List of ontologies") and, therefore, this technique is not vital for the calculation of similarity among concepts from different ontologies. As such, the total effort that I put into this project, with regard to my PhD, was relatively small when compared to other endeavours, since I recognised that the outputs of the project would not benefit multi-domain semantic similarity to a large extent.

Bibliography

[AB04] F. Azuaje and O. Bodenreider. "Incorporating ontology-driven similarity knowledge into functional genomics: an exploratory study". In: *IEEE Symposium on Bioinformatics and Bioengineering*. IEEE, 2004, pp. 317–324. ISBN: 0-7695-2173-8. DOI: 10.1109/BIBE.2004.1317360.

[AKK07] I. Astrova, N. Korda, and A. Kalja. "Storing OWL Ontologies in SQL3 Object-Relational Databases". In: *Engineering and Technology* 1.4 (2007), pp. 167–172. URL: www.wseas.us/e-library/conferences/2008/rhodes/aic/aic15.pdf.

[Alt90] S. F. Altschul et al. "Basic local alignment search tool." In: *Journal of Molecular Biology* 215.3 (1990), pp. 403–10. ISSN: 0022-2836. DOI: 10.1006/jmbi.1990.9999.

[Alt97] S. F. Altschul. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". In: *Nucleic Acids Research* 25.17 (Sept. 1997), pp. 3389–3402. ISSN: 13624962. DOI: 10.1093/nar/25.17.3389.

[Ash00] M. Ashburner et al. "Gene Ontology: tool for the unification of biology". In: *Nature Genetics* 25.1 (2000), pp. 25–29. URL: http://www.nature.com/ng/journal/v25/n1/abs/ng0500_25.html.

[Bar12] M. Barton. *Bioinformatics Career Survey*. 2012. URL: http://openwetware.org/wiki/Biogang:Projects/Bioinformatics_Career_Survey_2008 (visited on 11/26/2015).

[Bat02] A. Bateman et al. "The Pfam Protein Families Database". In: *Nucleic Acids Research* 30.1 (2002), pp. 276–280. ISSN: 13624962. DOI: 10.1093/nar/30.1.276.

[Bat12] D. S. Batista et al. "Toponym Disambiguation using Ontology-based Semantic Similarity". In: *Computational Processing of the Portuguese Language* 7243 (2012), pp. 179–185. DOI: 10.1007/978-3-642-28885-2_20.

[Bau11] C. Baumgartner et al. "Bioinformatic-driven search for metabolic biomarkers in disease." In: *Journal of Clinical Bioinformatics* 1.1 (2011), p. 2. ISSN: 2043-9113. DOI: 10.1186/2043-9113-1-2.

[Ber73] R. Berendzen. *Life beyond earth and the mind of man*. 1973. URL: http://hdl.handle.net/2060/19730022075.

[BHB09] C. Bizer, T. Heath, and T. Berners-Lee. "Linked Data - The Story So Far". In: *International Journal on Semantic Web and Information Systems* 5.3 (2009). Ed. by T. Heath, M. Hepp, and C. Bizer, pp. 1–22. ISSN: 15526283. DOI: 10.4018/jswis.2009081901.

[BHL01] T. Berners-Lee, J. Hendler, and O. Lassila. "The semantic web". In: *Scientific American* 284.5 (2001), pp. 28–37. URL: http://www.scientificamerican.com/article/the-semantic-web/.

[BHS05] F. Baader, I. Horrocks, and U. Sattler. "Description logics as ontology languages for the semantic web". In: *Mechanizing Mathematical Reasoning* (2005), pp. 1–21. URL: http://link.springer.com/chapter/10.1007/978-3-540-32254-2_14.

[Biz09] C. Bizer et al. "DBpedia - A crystallization point for the Web of Data". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 7.3 (Sept. 2009), pp. 154–165. ISSN: 15708268. DOI: 10.1016/j.websem.2009.07.002.

[BK98] P. Bork and E. V. Koonin. "Predicting functions from protein sequences–where are the bottlenecks?" In: *Nature Genetics* 18.4 (1998), pp. 313–318. ISSN: 1061-4036. DOI: 10.1038/ng0498-313.

[BN06] A. Ben-Hur and W. S. Noble. "Choosing negative examples for the prediction of protein-protein interactions." In: *BMC Bioinformatics* 7 Suppl 1 (2006), S2. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-S1-S2.

[Bol08] E. E. Bolton et al. "PubChem: integrated platform of small molecules and biological activities". In: *Annual Reports in Computational Chemistry*. Ed. by R. A. Wheeler and D. C. Spellmeyer. Vol. 4. American Chemical Society, Washington, DC, 2008 Apr, 2008. Chap. 12, pp. 217–241. ISBN: 9780444532503. DOI: 10.1016/S1574-1400(08)00012-1.

[Bud99] A. Budanitsky. "Lexical semantic relatedness and its application in natural language processing". PhD thesis. University of Toronto, 1999. URL: http://ukpmc.ac.uk/abstract/CIT/300791.

[Car04] J. J. Carroll et al. "Jena: Implementing the Semantic Web Recommendations". In: *International World Wide Web Conference*. 2004, pp. 74–83. ISBN: 1581139128. DOI: 10.1145/1013367.1013381.

[Car09] S. Carbon et al. "AmiGO: online access to ontology and annotation data." In: *Bioinformatics* 25.2 (Jan. 2009), pp. 288–9. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btn615.

[CAS09] I. F. Cruz, F. P. Antonelli, and C. Stroe. "AgreementMaker: efficient matching for large real-world schemas and ontologies". In: *VLDB Endowment*. Vol. 2. 2. VLDB Endowment, 2009, pp. 1586–1589. URL: http://dl.acm.org/citation.cfm?id=1687598.

[CL75] A. M. Collins and E. F. Loftus. "A spreading-activation theory of semantic processing". In: *Psychological Review* 82.6 (1975), pp. 407–428. URL: http://psycnet.apa.org/psycinfo/1976-03421-001.

[Cor04] S. de Coronado et al. "NCI Thesaurus: using science-based terminology to integrate cancer research results". In: *Studies in Health Technology and Informatics* 107.Pt 1 (2004), pp. 33–37. ISSN: 0926-9630. URL: http://www.ncbi.nlm.nih.gov/pubmed/15360769.

[Côt80] R. A. Côté. "Progress in Medical Information Management". In: *The Journal of the American Medical Association* 243.8 (Feb. 1980), p. 756. ISSN: 0098-7484. DOI: 10.1001/jama.1980.03300340032015.

[Cou12] F. M. Couto et al. "The Epidemic Marketplace Platform: towards semantic characterization of epidemiological resources using biomedical ontologies". In: *International Conference on Biomedical Ontologies*. 2012. URL: http://ceur-ws.org/Vol-897/demo_1.pdf.

[CP13] F. M. Couto and H. S. Pinto. "The next generation of similarity measures that fully explore the semantics in biomedical ontologies." In: *Journal of bioinformatics and computational biology* 11.5 (2013), p. 1371001. ISSN: 1757-6334. DOI: 10.1142/S0219720013710017.

[CR05] S. Chow and P. Rodgers. “Constructing area-proportional Venn and Euler diagrams with three circles”. In: *Euler Diagrams*. 2005, pp. 1–4. URL: https://kar.kent.ac.uk/14285/.

[CS11] F. M. Couto and M. J. Silva. “Disjunctive shared information between ontology concepts: application to Gene Ontology.” In: *Journal of Biomedical Semantics* 2.1 (Jan. 2011), p. 5. ISSN: 2041-1480. DOI: 10.1186/2041-1480-2-5.

[CSC07] F. M. Couto, M. J. Silva, and P. M. Coutinho. “Measuring semantic similarity between Gene Ontology terms”. In: *Data & Knowledge Engineering* 61.1 (Apr. 2007), pp. 137–152. ISSN: 0169023X. DOI: 10.1016/j.datak.2006.05.003.

[CSV05] J. C. Clemente, K. Satou, and G. Valiente. “Reconstruction of phylogenetic relationships from metabolic pathways based on the enzyme hierarchy and the gene ontology.” In: *International Conference on Genome Informatics*. Vol. 16. 2. Jan. 2005, pp. 45–55. URL: http://www.ncbi.nlm.nih.gov/pubmed/16901088.

[DB79] D. L. Davies and D. W. Bouldin. “A cluster separation measure.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1.2 (1979), pp. 224–227. ISSN: 0162-8828. DOI: 10.1109/TPAMI.1979.4766909.

[DCM12] DCMI Usage Board. *DCMI Metadata Terms*. Tech. rep. 2012. URL: http://dublincore.org/documents/dcmi-terms/.

[Deg08] K. Degtyarenko et al. “ChEBI: a database and ontology for chemical entities of biological interest”. In: *Nucleic Acids Research* 36.Database issue (2008), p. D344. DOI: 10.1093/nar/gkm791.

[DS05] A. Doms and M. Schroeder. “GoPubMed: exploring PubMed with the Gene Ontology.” In: *Nucleic Acids Research* 33.Web Server issue (2005), W783–6. ISSN: 1362-4962. DOI: 10.1093/nar/gki470.

[Ema14] E. Emadzadeh et al. “Unsupervised gene function extraction using semantic vectors.” In: *Database : The Journal of Biological Databases and Curation* 2014.i (2014), pp. 1–7. ISSN: 1758-0463. DOI: 10.1093/database/bau084.

[ES07] J. Euzenat and P. Shvaiko. *Ontology matching*. 1st. Springer-Verlang, 2007. ISBN: 3-540-49611-4. URL: http://book.ontologymatching.org/.

[FA15] W. W. M. Fleuren and W. Alkema. “Application of text mining in the biomedical domain.” In: *Methods (San Diego, Calif.)* 74 (2015), pp. 97–106. ISSN: 1095-9130. DOI: 10.1016/j.ymeth.2015.01.015.

[Faw04] T. Fawcett. “ROC graphs: Notes and practical considerations for researchers”. In: *Machine Learning* 31 (2004), pp. 1–38. DOI: 10.1.1.10.9777.

[Faw06] T. Fawcett. “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874. ISSN: 01678655. DOI: 10.1016/j.patrec.2005.10.010.

[FC10] J. D. Ferreira and F. M. Couto. “Semantic Similarity for Automatic Classification of Chemical Compounds”. In: *PLoS Computational Biology* 6.9 (Sept. 2010). Ed. by J. B. O. Mitchell, e1000937. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000937.

[FC11] J. D. Ferreira and F. M. Couto. “Generic semantic relatedness measure for biomedical ontologies”. In: *International Conference on Biomedical Ontologies*. 2011. URL: http://ceur-ws.org/Vol-833/paper16.pdf.

[Fer10] J. D. Ferreira et al. *The Geo-Net-PT / Yahoo ! GeoPlanet TM concordance*. Tech. rep. October. Departamento de Informática – Faculdade de Ciências - Universidade de Lisboa, 2010. DOI: DOI:10455/6677.

[Fer12] J. D. Ferreira et al. "Bringing epidemiology into the Semantic Web". In: *International Conference on Biomedical Ontologies*. 2012. URL: http://ceur-ws.org/Vol-897/session1-paper02.pdf.

[Fer13] J. D. Ferreira et al. "Digital preservation of epidemic resources: coupling metadata and ontologies". In: *International Conference on Preservation of Digital Objects*. Ed. by J. Borbinha, M. Nelson, and S. Knight. 2013. URL: http://xldb.di.fc.ul.pt/xldb/publications/Ferreira.etal:DigitalPreservationOf:2013_document.pdf.

[FGG97] N. Friedman, D. Geiger, and M. Goldszmidt. "Bayesian Network Classifiers". In: *Machine Learning* 29 (1997), pp. 131–163. ISSN: 0885-6125. DOI: 10.1023/A:1007465528199.

[FHC13] J. D. Ferreira, J. Hastings, and F. M. Couto. "Exploiting disjointness axioms to improve semantic similarity measures." In: *Bioinformatics* 29.21 (2013), pp. 2781–7. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt491.

[FM83] E. B. Fowlkes and C. L. Mallows. "A Method for Comparing Two Hierarchical Clusterings". In: *Journal of the American Statistical Association* 78.383 (1983), p. 553. ISSN: 01621459. DOI: 10.2307/2288117.

[FS99] C. V. Forst and K. Schulten. "Evolution of Metabolisms: A New Method for the Comparison of Metabolic Pathways Using Genomics Information". In: *Journal of Computational Biology* 6.3/4 (1999), pp. 343–360.

[GB12] V. N. Garla and C. Brandt. "Ontology-guided feature engineering for clinical text classification." In: *Journal of Biomedical Informatics* 45.5 (2012), pp. 992–8. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2012.04.010.

[GB14] R. Guha and D. Brickley. *RDF Schema 1.1*. Tech. rep. W3C, 2014. URL: http://www.w3.org/TR/2014/REC-rdf-schema-20140225/.

[GCA14] K. R. Gøeg, R. Cornet, and S. K. Andersen. "Clustering clinical models from local electronic health records based on semantic similarity". In: *Journal of Biomedical Informatics* 54 (2014), pp. 294–304. ISSN: 15320464. DOI: 10.1016/j.jbi.2014.12.015.

[Gen07] R. Gentleman. *Visualizing and distances using GO*. Tech. rep. 2007. URL: http://www.bioconductor.org/packages/release/bioc/vignettes/GOstats/inst/doc/GOvis.pdf.

[GGD13] C. Golbreich, J. Grosjean, and S. J. Darmoni. "The Foundational Model of Anatomy in OWL 2 and its use." In: *Artificial Intelligence in Medicine* 57.2 (2013), pp. 119–32. ISSN: 1873-2860. DOI: 10.1016/j.artmed.2012.11.002.

[Gre10] T. Grego et al. *Chemical and Metabolic Pathway Semantic Similarity*. Tech. rep. Department of Informatics, Faculty of Sciences, University of Lisbon, 2010. DOI: DOI:10455/3335.

[Gro12] A. Groß et al. "GOMMA results for OAEI 2012". In: *International Semantic Web Conference*. 2012. URL: http://disi.unitn.it/~p2p/OM-2012/oaei12_paper3.pdf.

[Gru93] T. R. Gruber. "A translation approach to portable ontology specifications". In: *Knowledge Acquisition* 5.2 (1993), pp. 199–220. ISSN: 1042-8143. DOI: 10.1.1.101.7493.

[GSG04] P. Grenon, B. Smith, and L. J. Goldberg. "Biodynamic ontology: Applying BFO in the biomedical domain". In: *Studies in Health Technology and Informatics* 102.ii (2004), pp. 20–38. ISSN: 09269630. DOI: 10.3233/978-1-60750-945-5-20.

[Gua98] N. Guarino. "Formal Ontology and Information Systems". In: *Proceedings of the first international conference*. Vol. 46. June. 1998, pp. 3–15. ISBN: 9051993994. DOI: 10.1.1.29.1776.

[Guo06] X. Guo et al. "Assessing semantic similarity measures for the characterization of human regulatory pathways." In: *Bioinformatics* 22.8 (Apr. 2006), pp. 967–73. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl042.

[GVC13] P. H. Guzzi, P. Veltri, and M. Cannataro. "OntoPIN: an ontology-annotated PPI database." In: *Interdisciplinary Sciences, Computational Life Sciences* 5.3 (2013), pp. 187–95. ISSN: 1867-1462. DOI: 10.1007/s12539-013-0173-x.

[GZB06] C. Golbreich, S. Zhang, and O. Bodenreider. "The foundational model of anatomy in OWL: Experience and perspectives". In: *Web Semantics* 4.3 (2006), pp. 181–195. ISSN: 16130073. DOI: 10.1016/j.websem.2006.05.007. eprint: NIHMS150003.

[Har14] S. Harispe et al. "A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain". In: *Journal of Biomedical Informatics* 48.November (Nov. 2014), pp. 38–53. ISSN: 15320464. DOI: 10.1016/j.jbi.2013.11.006.

[Has12] J. Hastings et al. "Modular Extensions to the ChEBI Ontology". In: *International Conference on Biomedical Ontologies*. 2012. URL: http://ceur-ws.org/Vol-897/poster_7.pdf.

[Has13] J. Hastings et al. "The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013." In: *Nucleic Acids Research* 41.Database issue (Jan. 2013), pp. D456–63. ISSN: 1362-4962. DOI: 10.1093/nar/gks1146.

[HB11] M. Horridge and S. Bechhofer. "The OWL API: A Java API for OWL ontologies". In: *Semantic Web* 2.1 (2011), pp. 11–21. ISSN: 1570-0844. DOI: 10.3233/SW-2011-0025.

[HDG13] R. Hoehndorf, M. Dumontier, and G. V. Gkoutos. "Evaluation of research in biomedical ontologies". In: *Briefings in Bioinformatics* 14.6 (2013), pp. 696–712. ISSN: 14675463. DOI: 10.1093/bib/bbs053.

[Hen09] J. Henß et al. "A database backend for OWL". In: *OWL: Experiences and Directions* 2009.5th Int. Workshop on OWL: Experiences and Directions Owled (2009). URL: http://ceur-ws.org/Vol-529/owled2009_submission_3.pdf.

[HF64] C. Hansch and T. Fujita. "ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure". In: *Journal of the American Chemical Society* 86.8 (1964), pp. 1616–1626. ISSN: 0002-7863. DOI: 10.1021/ja01062a035.

[HK11] J. Hauke and T. Kossowski. "Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data". In: *Quaestiones Geographicae* 30.2 (2011), pp. 87–93. ISSN: 0137-477X. DOI: 10.2478/v10117-011-0021-1.

[HS03] M. Heymans and A. Singh. "Deriving phylogenetic trees from the similarity analysis of metabolic pathways". In: *Bioinformatics* 19.Suppl 1 (2003), p. i138. URL: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/19/suppl_1/i138.

[HS13] S. Harris and A. Seaborne. *SPARQL 1.1 Query Language*. Tech. rep. W3C, 2013. DOI: citeulike-article-id:2620569.

[Hu12] Y. Hu et al. “Integrating various resources for gene name normalization.” In: *PloS One* 7.9 (2012), e43558. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0043558.

[JB10] S. Jain and G. D. Bader. “An improved method for scoring protein-protein interactions using semantic similarity within the Gene Ontology”. In: *BMC Bioinformatics* 11 (Jan. 2010), p. 562. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-562.

[JC97] J. J. Jiang and D. W. Conrath. “Semantic similarity based on corpus statistics and lexical taxonomy”. In: *Research on Computational Linguistics*. 1997. URL: http://www.aclclp.org.tw/rocling/1997/M02.pdf.

[Jut15] N. Juty et al. “BioModels: Content, Features, Functionality, and Use.” In: *CPT: Pharmacometrics & Systems Pharmacology* 4.2 (Feb. 2015), e3. ISSN: 2163-8306. DOI: 10.1002/psp4.3.

[KAM94] G. Klebe, U. Abraham, and T. Mietzner. “Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity”. In: *Journal of Medicinal Chemistry* 37.24 (1994), pp. 4130–4146. ISSN: 00222623. DOI: 10.1021/jm00050a010.

[KKK13] J. Kim, J.-S. Kwon, and S. Kim. “Gene Set Analyses of Genome-Wide Association Studies on 49 Quantitative Traits Measured in a Single Genetic Epidemiology Dataset.” In: *Genomics & Informatics* 11.3 (2013), pp. 135–141. ISSN: 1598-866X. DOI: 10.5808/GI.2013.11.3.135.

[KMP04] E. P. Klement, R. Mesiar, and E. Pap. “Triangular norms. Position paper I: basic analytical and algebraic properties”. In: *Fuzzy Sets and Systems* 143.1 (Apr. 2004), pp. 5–26. ISSN: 01650114. DOI: 10.1016/j.fss.2003.06.007.

[Köh09] S. Köhler et al. “Clinical diagnostics in human genetics with semantic similarity searches in ontologies.” In: *American Journal of Human Genetics* 85.4 (Oct. 2009), pp. 457–64. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2009.09.003.

[KTM15] M. R. Kamdar, T. Tudorache, and M. A. Musen. “Investigating Term Reuse and Overlap in Biomedical Ontologies”. In: *International Conference on Biomedical Ontologies*. 2015. URL: http://ceur-ws.org/Vol-1515/regular9.pdf.

[LD06] Z. Lei and Y. Dai. “Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction.” In: *BMC Bioinformatics* 7.1 (Jan. 2006), p. 491. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-491.

[LFC14] A. Lamurias, J. D. Ferreira, and F. M. Couto. “Identifying interactions between chemical entities in biomedical text”. In: *Journal of Interactive Bioinformatics (JIB)* (2014), pp. 1–18. ISSN: 1613-4516. DOI: 10.2390/biecoll-jib-2014-247.

[LFC15] A. Lamurias, J. D. Ferreira, and F. M. Couto. “Improving chemical entity recognition through h-index based semantic similarity.” In: *Journal of Cheminformatics* 7.Suppl 1 Text mining for chemistry and the CHEMDNER track (Jan. 2015), S13. ISSN: 1758-2946. DOI: 10.1186/1758-2946-7-S1-S13.

[LI10] P. O. Larsen and M. von Ins. “The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index.” In: *Scientometrics* 84.3 (Sept. 2010), pp. 575–603. ISSN: 0138-9130. DOI: 10.1007/s11192-010-0202-z.

[Li10a] C. Li et al. “BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models”. In: *BMC Systems Biology* 4.92 (2010). DOI: doi:10.1186/1752-0509-4-92.

[Li10b] X. Li et al. “Computational approaches for detecting protein complexes from protein interaction networks: a survey.” In: *BMC genomics* 11 Suppl 1 (2010), S3. ISSN: 1471-2164. DOI: 10.1186/1471-2164-11-S1-S3.

[Li11] J. Li et al. “DOSim: An R package for similarity between diseases based on Disease Ontology.” In: *BMC Bioinformatics* 12.1 (June 2011), p. 266. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-266.

[Lie08] T. Liebig et al. “OWLlink: DIG for OWL 2”. In: *OWL: Experiences and Directions*. Vol. 432. 2008. URL: http://ceur-ws.org/Vol-432/owled2008eu_submission_26.pdf.

[Lin98] D. Lin. “An information-theoretic definition of similarity”. In: *International Conference on Machine Learning*. Vol. 1. 1998, pp. 296–304. URL: http://webdocs.cs.ualberta.ca/~lindek/papers/sim.pdf.

[LO12] P. Lopes and J. L. Oliveira. “COEUS: "semantic web in a box" for biomedical applications.” In: *Journal of Biomedical Semantics* 3.1 (Dec. 2012), p. 11. ISSN: 2041-1480. DOI: 10.1186/2041-1480-3-11.

[Lop10] L. F. Lopes et al. “Epidemic Marketplace: an information management system for epidemiological data”. In: *Information Technology in Bio- and Medical-Informatics*. 2010, pp. 31–44. URL: http://link.springer.com/chapter/10.1007/978-3-642-15020-3_3.

[Lor03] P. W. Lord et al. “Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation”. In: *Bioinformatics* 19.10 (July 2003), pp. 1275–1283. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btg153.

[LT12] K. Lehmann and A. Y. Turhan. “A Framework for Semantic-based Similarity Measures for ELH-Concepts”. In: *Logics in Artificial Intelligence* (2012), pp. 307–319. URL: http://link.springer.com/chapter/10.1007/978-3-642-33353-8_24.

[McG02] D. L. McGuinness. “Ontologies come of age”. In: *Spinning the Semantic Web*. Ed. by D. Fensel et al. The MIT Press, 2002. Chap. 6, pp. 171–192. ISBN: 0262062321. DOI: 10.1.1.546.224.

[McI11] B. T. McInnes et al. “Knowledge-based Method for Determining the Meaning of Ambiguous Biomedical Terms Using Information Content Measures of Similarity”. In: *AMIA Annual Symposium Proceedings* 2011 (2011), p. 895. ISSN: 1942-597X. URL: http://www.ncbi.nlm.nih.gov/pubmed/22195148.

[MD12] S. Mathur and D. Dinakarpandian. “Finding disease similarity based on implicit semantic similarity”. In: *Journal of Biomedical Informatics* 45.2 (2012), pp. 363–71. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2011.11.017.

[Mil95] G. A. Miller. “WordNet: a lexical database for English”. In: *Communications of the ACM* 38.11 (Nov. 1995), pp. 39–41. ISSN: 00010782. DOI: 10.1145/219717.219748.

[MN09] H. Al-Mubaid and H. A. Nguyen. “Measuring Semantic Similarity Between Biomedical Concepts Within Multiple Ontologies”. In: *IEEE Transactions on Systems, Man, and Cybernetics* 39.4 (July 2009), pp. 389–398. ISSN: 1094-6977. DOI: 10.1109/TSMCC.2009.2020689.

[Mos15] G. P. Moss. *Enzyme Nomenclature*. 2015. URL: http://www.chem.qmul.ac.uk/iubmb/enzyme/.

[Mot12] B. Motik et al. *OWL 2 Web Ontology Language - Structural Specification and Functional-Style Syntax (Second Edition)*. Tech. rep. 2012. URL: http://www.w3.org/TR/2012/REC-owl2-syntax-20121211/.

[MPG12] B. Motik, P. F. Patel-Schneider, and B. C. Grau. *OWL 2 Web Ontology Language Direct Semantics (Second Edition)*. Tech. rep. W3C, 2012. URL: http://www.w3.org/TR/2012/REC-owl2-direct-semantics-20121211/.

[Mun11] C. J. Mungall et al. "Cross-product extensions of the Gene Ontology". In: *Journal of Biomedical Informatics* 44.1 (2011), pp. 80–6. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2010.02.002.

[Nal06] R. Nalichowski et al. "Calculating the benefits of a Research Patient Data Repository." In: *AMIA Annual Symposium Proceedings* (Jan. 2006), p. 1044. ISSN: 1942-597X. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839563/.

[NB03] D. Nardi and R. J. Brachman. "An Introduction to Description Logics". In: *The Description Logic Handbook: Theory, Implementation and Applications*. Ed. by F. Baader. 2nd ed. Cambridge University Press, 2003. Chap. 1, pp. 5–41. ISBN: 0-521-78176-0. DOI: 10.1017/CBO9780511711787.003.

[NM01] N. F. Noy and D. L. McGuinness. "Ontology development 101: A guide to creating your first ontology". In: *Development* 32 (2001), pp. 1–25. ISSN: 09333657. DOI: 10.1016/j.artmed.2004.01.014.

[Noy09] N. F. Noy et al. "BioPortal: ontologies and integrated data resources at the click of a mouse." In: *Nucleic Acids Research* 37.Web Server issue (July 2009), W170–3. ISSN: 1362-4962. DOI: 10.1093/nar/gkp440.

[NR02] R. Nair and B. Rost. "Sequence conserved for subcellular localization." In: *Protein Science* 11 (2002), pp. 2836–2847. ISSN: 0961-8368. DOI: 10.1110/ps.0207402.

[OB71] B. M. Oliver and J. Billingham. "Project Cyclops: A Design Study of a System for Detecting Extraterrestrial Intelligent Life". In: *The 1971 NASA/ASEE Summer Facultu Fellowship Program* (1971). URL: https://seti.berkeley.edu/sites/default/files/19730010095_1973010095.pdf.

[OD09] M. J. O'Connor and A. Das. "SQWRL: A query language for OWL". In: *OWL: Experiences and Directions* 529.Owled (2009). ISSN: 16130073. URL: http://ceur-ws.org/Vol-529/owled2009_submission_42.pdf.

[Oga99] H. Ogata et al. "KEGG: Kyoto Encyclopedia of Genes and Genomes". In: *Nucleic Acids Research* 27.1 (1999), pp. 29–34. ISSN: 0305-1048. DOI: 10.1093/nar/27.1.29.

[Osb09] J. D. Osborne et al. "Annotating the human genome with Disease Ontology". In: *BMC Genomics* 10.Suppl 1 (Jan. 2009), S6. ISSN: 1471-2164. DOI: 10.1186/1471-2164-10-S1-S6.

[Ove13] W. F. Overton. *Reasoning, Necessity, and Logic: Developmental Perspectives*. Psychology Press, 2013, p. 344. ISBN: 1134735146. URL: https://books.google.com/books?hl=en&lr=&id=Y2gTxnlDY20C&pgis=1.

[Pap03] J. A. Papin et al. "Metabolic pathways in the post-genome era." In: *Trends in Biochemical Sciences* 28.5 (2003), pp. 250–258. ISSN: 0968-0004. DOI: 10.1016/S0968-0004(03)00064-1.

[Ped07] T. Pedersen et al. “Measures of semantic similarity and relatedness in the biomedical domain.” In: *Journal of Biomedical Informatics* 40.3 (June 2007), pp. 288–99. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2006.06.004.

[Pes08] C. Pesquita et al. “Metrics for GO based protein semantic similarity: a systematic evaluation.” In: *BMC Bioinformatics* 9 Suppl 5 (2008), S4. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-S5-S4.

[Pes09a] C. Pesquita et al. “CESSM: Collaborative Evaluation of Semantic Similarity Measures”. In: *JB2009: Challenges in Bioinformatics*. 2009. URL: http://homepages.di.fc.ul.pt/~fjmc/files/workshop%20cpesquita-jb2009.pdf.

[Pes09b] C. Pesquita et al. “Semantic similarity in biomedical ontologies”. In: *PLoS computational biology* 5.7 (2009), e1000443. DOI: 10.1371/journal.pcbi.1000443.

[Por08] M. Porta, ed. *A dictionary of epidemiology*. 5th. Oxford University Press, 2008, 2008. ISBN: 9780199976737. URL: https://global.oup.com/academic/product/a-dictionary-of-epidemiology-9780199976737.

[Qui68] M. Quillian. “Semantic memory”. In: *Semantic Information Processing* (1968). Ed. by M. Minsky, pp. 227–270.

[Rad89] R. Rada et al. “Development and application of a metric on semantic nets”. In: *IEEE Transactions on Systems, Man, and Cybernetics* 19.1 (1989), pp. 17–30. ISSN: 00189472. DOI: 10.1109/21.24528.

[RE03] M. A. Rodríguez and M. J. Egenhofer. “Determining Semantic Similarity among Entity Classes from Different Ontologies”. In: *Knowledge Creation Diffusion Utilization* 15.2 (2003), pp. 442–456. DOI: 10.1109/TKDE.2003.1185844.

[Res95] P. Resnik. “Using information content to evaluate semantic similarity in a taxonomy”. In: *International Joint Conference on Artificial Intelligence*. Vol. 1. 1995. eprint: 9511007v1. URL: http://dl.acm.org/citation.cfm?id=1625914.

[Res99] P. Resnik. “Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language”. In: *Journal of Artificial Intelligence Research* 11 (1999), pp. 95–130. DOI: 10.1613/jair.514. eprint: 1105.5444.

[RM03] C. Rosse and J. L. V. Mejino. “A reference ontology for biomedical informatics: the Foundational Model of Anatomy”. In: *Journal of Biomedical Informatics* 36.6 (2003), pp. 478–500. URL: http://linkinghub.elsevier.com/retrieve/pii/S1532046403001278.

[Rog63] F. B. Rogers. “Medical subject headings.” In: *Bulletin of the Medical Library Association* 51 (Jan. 1963), pp. 114–6. ISSN: 0025-7338. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC197951/.

[Ros10] B. Rosner. *Fundamentals of biostatistics*. 7th. Cengage Learning, 2010. ISBN: 0538733497.

[Ros95] C. Rosse et al. “Enhancements of anatomical information in UMLS knowledge sources”. In: *Computer Application in Medical Care*. 1995, pp. 873–7. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2579218/.

[Rue08] A. Ruepp et al. “CORUM: the comprehensive resource of mammalian protein complexes.” In: *Nucleic Acids Research* 36.Database issue (2008), pp. D646–50. ISSN: 1362-4962. DOI: 10.1093/nar/gkm936.

[Sal04] L. Salwinski et al. “The Database of Interacting Proteins: 2004 update.” In: *Nucleic Acids Research* 32.Database issue (2004), pp. D449–D451. ISSN: 1362-4962. DOI: 10.1093/nar/gkh086.

[Sán12a] D. Sánchez et al. “Enabling semantic similarity estimation across multiple ontologies: an evaluation in the biomedical domain.” In: *Journal of Biomedical Informatics* 45.1 (Feb. 2012), pp. 141–155. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2011.10.005.

[Sán12b] D. Sánchez et al. “Ontology-based semantic similarity: A new feature-based approach”. In: *Expert Systems with Applications* 39.9 (July 2012), pp. 7718–7728. ISSN: 09574174. DOI: 10.1016/j.eswa.2012.01.082.

[SB13] D. Sánchez and M. Batet. “A semantic similarity method based on information content exploiting multiple ontologies”. In: *Expert Systems with Applications* 40.4 (Aug. 2013), pp. 1393–1399. ISSN: 09574174. DOI: 10.1016/j.eswa.2012.08.049.

[SBH06] N. Shadbolt, T. Berners-Lee, and W. Hall. “The Semantic Web Revisited”. In: *IEEE Intelligent Systems* 21.3 (May 2006), pp. 96–101. ISSN: 1541-1672. DOI: 10.1109/MIS.2006.62.

[SBI11] D. Sánchez, M. Batet, and D. Isern. “Ontology-based information content computation”. In: *Knowledge-Based Systems* 24.2 (Mar. 2011), pp. 297–303. ISSN: 09507051. DOI: 10.1016/j.knosys.2010.10.001.

[SE05] P. Shvaiko and J. Euzenat. “A survey of schema-based matching approaches”. In: *Journal on Data Semantics* IV (2005), pp. 146–171. URL: http://link.springer.com/chapter/10.1007/11603412_5.

[Sek15] M. K. Sekhwal et al. “Identification of drought-induced transcription factors in Sorghum bicolor using GO term semantic similarity.” In: *Cellular & Molecular Biology Letters* 20.1 (2015), pp. 1–23. ISSN: 1689-1392. DOI: 10.2478/s11658-014-0223-3.

[Sin31] C. J. Singer. *A Short History of Biology*. Oxford University Press, 1931.

[SKL12] M. Schulz, E. Klipp, and W. Liebermeister. “Propagating semantic information in biochemical network models.” In: *BMC Bioinformatics* 13.1 (Jan. 2012), p. 18. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-18.

[SLA10] A. Schlicker, T. Lengauer, and M. Albrecht. “Improving disease gene prioritization using the semantic similarity of Gene Ontology terms”. In: *Bioinformatics* 26.18 (Sept. 2010), pp. i561–i567. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq384.

[Smi07] B. Smith et al. “The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.” In: *Nature Biotechnology* 25.11 (Nov. 2007), pp. 1251–5. ISSN: 1087-0156. DOI: 10.1038/nbt1346.

[SP07] E. Sirin and B. Parsia. “SPARQL-DL: SPARQL query for OWL-DL”. In: *OWL: Experiences and Directions*. Vol. 258. 2007, pp. 1–10. URL: http://ceur-ws.org/Vol-258/paper14.pdf.

[Spa05] I. Spasic et al. “Text mining and ontologies in biomedicine: making sense of raw text.” In: *Briefings in Bioinformatics* 6.3 (Sept. 2005), pp. 239–51. ISSN: 1467-5463. URL: http://www.ncbi.nlm.nih.gov/pubmed/16212772.

[SSL13] U. Sattler, R. D. Stevens, and P. W. Lord. “(I can’t get no) satisfiability”. In: *Ontogenesis* (May 2013). URL: http://ontogenesis.knowledgeblog.org/1329.

[Sub05] A. Subramanian et al. “Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide”. In: *Proceedings of the National Academy of Sciences*. Vol. 102. 43. 2005, pp. 15545–15550. DOI: 10.1073/pnas.0506580102.

[Sup11] F. Supek et al. “REVIGO summarizes and visualizes long lists of gene ontology terms.” In: *PloS One* 6.7 (2011), e21800. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0021800.

[SVH04] N. Seco, T. Veale, and J. Hayes. “An intrinsic information content metric for semantic similarity in WordNet”. In: *ECAI* 16 (2004), p. 1089. URL: http://afflatus.ucd.ie/papers/ecai2004b.pdf.

[SW81] T. F. Smith and M. S. Waterman. “Identification of common molecular subsequences.” In: *Journal of Molecular Biology* 147.1 (1981), pp. 195–197. ISSN: 00222836. DOI: 10.1016/0022-2836(81)90087-5.

[Swe74] P. T. Swender. “Computer-Assisted Diagnosis”. In: *Archives of Pediatrics & Adolescent Medicine* 127.6 (June 1974), p. 859. ISSN: 1072-4710. DOI: 10.1001/archpedi.1974.02110250085012.

[Tan14] F. Tan et al. “Drug repositioning by applying ’expression profiles’ generated by integrating chemical structure similarity and gene semantic similarity.” In: *Molecular BioSystems* 10.5 (2014), pp. 1126–38. ISSN: 1742-2051. DOI: 10.1039/c3mb70554d.

[Tao07] Y. Tao et al. “Information theory applied to the sparse gene ontology annotation network to predict novel gene function.” In: *Bioinformatics* 23.13 (2007), pp. i529–38. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btm195.

[The10] The UniProt Consortium. “The Universal Protein Resource (UniProt) in 2010.” In: *Nucleic Acids Research* 38.Database issue (Jan. 2010), pp. D142–8. ISSN: 1362-4962. DOI: 10.1093/nar/gkp846.

[The12] The Gene Ontology Consortium. *An Introduction to the Gene Ontology*. 2012. URL: http://www.geneontology.org/GO.doc.shtml (visited on 10/28/2015).

[Tve77] A. Tversky. “Features of similarity”. In: *Psychological Review* 84.4 (1977), pp. 327–352. URL: http://psycnet.apa.org/journals/rev/84/4/327/.

[Vaf13] F. Vafaee et al. “Novel semantic similarity measure improves an integrative approach to predicting gene functional associations.” In: *BMC Systems Biology* 7 (Jan. 2013), p. 22. ISSN: 1752-0509. DOI: 10.1186/1752-0509-7-22.

[Var05] G. Varelas et al. “Semantic similarity methods in wordNet and their application to information retrieval on the web”. In: *International Workshop on Web Information and Data Management*. 2005, pp. 10–16. ISBN: 1595931945. DOI: 10.1145/1097047.1097051.

[Vis11] T. Vision et al. “Similarity between semantic description sets: addressing needs beyond data integration”. In: *International Semantic Web Conference*. 2011. URL: http://ceur-ws.org/Vol-783/paper8.pdf.

[Wan04] H. Wang et al. “Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships”. In: *International Geoscience and Remote Sensing*. Vol. 2004. IEEE, Oct. 2004, pp. 25–31. ISBN: 0-7803-8728-7. DOI: 10.1109/CIBCB.2004.1393927.

[Wan07] J. Z. Wang et al. “A new method to measure the semantic similarity of GO terms”. In: *Bioinformatics* 23.10 (May 2007), pp. 1274–1281. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btm087.

[Weh08] M. Wehling. "Translational medicine: science or wishful thinking?" In: *Journal of Translational Medicine* 6 (Jan. 2008), p. 31. ISSN: 1479-5876. DOI: 10.1186/1479-5876-6-31.

[Whe11] P. L. Whetzel et al. "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications." In: *Nucleic Acids Research* 39.Web Server issue (July 2011), W541–5. ISSN: 1362-4962. DOI: 10.1093/nar/gkr469.

[Wil08] A. J. Williams. "A perspective of publicly accessible/open-access chemistry databases." In: *Drug Discovery Today* 13.11-12 (June 2008), pp. 495–501. ISSN: 1359-6446. DOI: 10.1016/j.drudis.2008.03.017.

[Wil12] A. J. Williams et al. "Open PHACTS: Semantic interoperability for drug discovery". In: *Drug Discovery Today* 17.21-22 (2012), pp. 1188–1198. ISSN: 13596446. DOI: 10.1016/j.drudis.2012.05.016.

[WJB04] S. K. Wyman, R. K. Jansen, and J. L. Boore. "Automatic annotation of organellar genomes with DOGMA". In: *Bioinformatics* 20.17 (2004), pp. 3252–3255. ISSN: 13674803. DOI: 10.1093/bioinformatics/bth352.

[WL03] J. C. Whisstock and A. M. Lesk. "Prediction of protein function from protein sequence and structure." In: *Quarterly Reviews of Biophysics* 36.3 (2003), pp. 307–340. ISSN: 0033-5835. DOI: 10.1017/S0033583503003901.

[WLT05] J. D. Watson, R. A. Laskowski, and J. M. Thornton. "Predicting protein function from sequence and structural data". In: *Current Opinion in Structural Biology* 15 (2005), pp. 275–284. ISSN: 0959-440X (PRINT). DOI: doi:10.1016/j.sbi.2005.04.003.

[Wol76] D. A. Wolfe. "On testing equality of related correlation coefficients". In: *Biometrika* 63.1 (Jan. 1976), pp. 214–215. ISSN: 0006-3444. DOI: 10.1093/biomet/63.1.214.

[WP94] Z. Wu and M. Palmer. "Verbs semantics and lexical selection". In: *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* -. Association for Computational Linguistics, 1994, pp. 133–138. DOI: 10.3115/981732.981751.

[Wu06] X. Wu et al. "Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations." In: *Nucleic Acids Research* 34.7 (2006), pp. 2137–50. ISSN: 1362-4962. DOI: 10.1093/nar/gkl219.

[Wu13] X. Wu et al. "Improving the Measurement of Semantic Similarity between Gene Ontology Terms and Gene Products: Insights from an Edge- and IC-Based Hybrid Method." In: *PloS One* 8.5 (Jan. 2013), e66745. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0066745.

[Xia11] Z. Xiang et al. "Ontobee: A linked data server and browser for ontology terms". In: *International Conference on Biomedical Ontologies*. Vol. 833. 2011, pp. 279–281. URL: http://ceur-ws.org/Vol-833/paper48.pdf.

[Xu13] Y. Xu et al. "A novel insight into Gene Ontology semantic similarity". In: *Genomics* 101.6 (2013), pp. 368–375. ISSN: 08887543. DOI: 10.1016/j.ygeno.2013.04.010.

[YNP12] H. Yang, T. Nepusz, and A. Paccanaro. "Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty." In: *Bioinformatics* 28.10 (May 2012), pp. 1383–9. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts129.

[Zho06] J. Zhou et al. “Minerva : A Scalable OWL Ontology Storage and Inference System”. In: *Asian Semantic Web Conference*. Dl. 2006, pp. 429–443. ISBN: 0302-9743. DOI: 10.1007/11836025_42.

[ZK14] N. Zina and N. Kaouther. “Automatically building database from biomedical ontology”. In: *International Work-Conference on Bioinformatics and Biomedical Engineering*. 2014, pp. 1403–1411. URL: http://iwbbio.ugr.es/2014/papers/IWBBIO_2014_paper_147.pdf.

[ZZ07] M. L. Zhang and Z. H. Zhou. “ML-KNN: A lazy learning approach to multi-label learning”. In: *Pattern Recognition Letters* 40.7 (2007), pp. 2038–2048. ISSN: 00313203. DOI: 10.1016/j.patcog.2006.12.019.